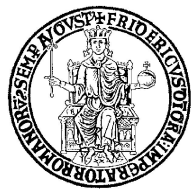


UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"
DIPARTIMENTO DI MATEMATICA E STATISTICA

DOTTORATO IN STATISTICA
XX CICLO

ADVANCED TOOLS FOR SOCIAL NETWORK ANALYSIS: COMPLEX NETWORK MODELS AND SPECTRAL GRAPH THEORY

Domenico De Stefano

ANNO ACCADEMICO 2007 - 2008

Contents

List of Figures

List of Tables

Introduction

In a 1983 paper Frank Harary, the father of the modern graph theory and one of the most prolific ”network scientist”, stated:

> [...] relations between graph theoretic concepts, as embodied in theorems, are seldom seen in network analysis [...] network analysts use only the concepts and terminology of graph theory while ignoring its theorems [...] The dynamic power of graph theory lies not in its terminology but, like any other branch of mathematics, in its theorems [...] Network analysts [...] make too little use of *theory* of graph [47].

More than twenty years have passed since the works of Harary and of the other earliest network scientists. Nowadays the analysis of network structure is currently one of the most interesting fields in several disciplines: from modern physics to social statistics, from biology to sociology. Especially social network analysis (SNA) has become one of the most explored field in applied statistics and social science methodology. Indeed, thanks to the emergent pervasive ”real structures” as the web communities, in a most theoretical fashion we assist in social sciences to the (re-)discovery of the so-called network paradigm.

A ever increasing number of studies, publications and manuscripts, but also pop-culture manifestations of interest (for example the famous ”Kevin Bacon Game” or the idea of ”six degrees of separation”), are dedicated to the social networks. A well known study on the trend of the publications in which the topic ”social networks” is treated, showed that since the 1965, the year of the Milgram’s small world problem, until 2005 the number of scientific works in which the words ”social” and ”networks” appear, in

both titles and abstracts, has grown exponentially [**?**].

SNA represents now the main statistical technique for the analysis of the *interactional component* of social behaviour. It is a fomal way to interpret social phenomena that can furnish to social sciences, and in particular to sociology, the long time missing scientific tool to interpret the society. However, since its youngness and multidisciplinarity, SNA still needs to strongly define the foundations of its theoretical framework and of its analytical tools.

In this thesis we suggest to adopt some theoretical and analytical advances. In particular we propose: i) to identify a theoretical framework useful to include SNA in the larger context of the network science; ii) to extend the SNA analytical tools exploring the advantages deriving from the use of more advanced graph theory results such as spectral decomposition of graphs.

For the first point we suggest to introduce in SNA a more ”systemic” point of view. In other words, to consider a social network as a representation of an ”ensemble” (i.e. the society) in which the parts interact independently from the global system but, at the same time, they determine the state of the entire set. This without forgotting that system and its part are strongly interrelated. Indeed, single part behaviour may influence the whole social network as well as the whole system may influence the subjective choices. For example, let imagine that a single individual interact in an highly clustered social structure. If he acts as single in such a topological relational space, the probability of becoming an isolated individual is very high. Let us think instead that the actor's behaviour is modelled according to the system structure: the probability of obtaining links, consensus or other ”relational resources” now becomes higher. In this sense, the system topology influence the individual's behaviour. On the other hand, the collective dynamics may change the topology of the social system if they radically differentiate the way in which links are created or deactivated.

With regard to the analytical advances, a possible direction is suggested directly from Frank Harary in our opening statement. In SNA a deeper investigation of what is done in advanced graph theory could lead toward extreme useful experimentation. The Harary's suggestion is still actual. Often, even nowadays, the social analysts forgot

about the fact that the most important formalisation in network analysis is represented by the theory of the graph structures. Sometimes social analysts behave more as social scientists that as statisticians or quantitative scientists. For example let consider the spectral graph theory (see chapter 3 on page 39) which represents the ”link” between graph heory, from one hand, and linear algebra, on the other hand. If we think that the most recent advances in statistical analysis are derived by considering data dipped in vector spaces we could legitimate the adoption of this special mathematical topic in the analysis of social network. We definitely think that the connection between linear algebra, graph theory and SNA could open the door that allows to use advanced multivariate statistical analysis in networks study.

If SNA, as a branch of statistics, can obtain advantages by using this kind of tools, we believe that also the statistical methodology *tout court* can be enriched by this tight connection between formalized graphs and linear algebra. In a recent work Ludovic Lebart[1] illustrated the advantages derived by taking into account the *graph structure* of the data. In particular he stated:

> [...] when dealing with contingency tables, theoretical links can be established in some cases in which a particular clustering technique provides hierarchical indexes that coincide with the eigenvalues from the correspondence analysis performed on the same table. [...] In a quite different context, the introduction of local metrics leads to a series of hybrid methods resembling projection pursuit algorithms that enrich the visualizations of clusters. *More recently, clustering techniques using the laplacians of graphs suggest new approaches*. [...] the graph structures defined by either the series of nearest neighbors or thresholds of distances appear to have interesting spectral properties. They may provide a valuable bridge between these two large families[2] [**?**].

In the same way we think that the ”use of the laplacians of graphs could suggest new approaches” and useful developments in SNA. We exepect that the experimentation of

[1] We refer to the invited lecture of L. Lebart during the 7^{th} *International Conference on Social Sciences and Methodology RC 33 - Logic and Methodology in Sociology* took place at University of Naples ”Federico II” in September 2008.

[2] Lebart spoke about principal axes analysis and clustering techniques.

new perspectives should represent the principal characteristic of an interdisciplinary field of study as SNA. To summarize, as it has already emerged from the previous discussion, the present manuscript is based on two main motivations:

- the beliefs of the author that social network theory should adopt a well denfined theoretical framework that allows to implement results and experimentation coming up from that new discipline known as network science.
- the fact that very little of graph theory tools are actually used in SNA, especially the useful relation between linear algebra and graph theory, implemented in the so-called spectral graph theory;

In particular, the former point is widely explored throughout this dissertation by mean of the study of the analogies and the intimate interrelation between the social network analysis and the complex network theory, presented in related chapters. We believe that, as part of a multidisciplinary field of study, it would be worth to implement, in SNA, the tools developed in network science and, at the same time, to look for new advances, directly from graph theory. Operatively, in order to use the approach of the general network theory in SNA, it is essential to consider social networks as special case of complex networks [1]. This analogy is not at all a matter of reductionism, but it is the basis of the systemic point of view we mentioned in the beginning. By means of it, it would be possible to build in the corpus of social networks methodology, the models and the laws that govern the complex networks.

To summarize, we will traduce these two motivations in two topics: first, the observation that the social networks could be viewed, with more advantages, as a particular specification of a complex network allowing, consequently, the possibilities of using complex networks models in the study of social networks; second, to show how the new advanced tools of spectral graph theory and complex networks models, adopted in SNA can, from one hand, solve specific problems and, on the other hand, provides the analytical way to open new branches of research. The double path on which we will move throughout this work converges on a single one: the specification of a new way to think of social networks, and the purpose to include SNA in a more organic place within the network science. The present work is divided in seven chapters. In particular

in the *chapter* 1, we introduce the basic concepts of the network perspective and of the ”network science”, starting from its history and the evolution of the relational approach in several disciplines. We conclude the chapter defining social networks in terms of complex system and discussing the applicability of the advanced tools of graph theory to the SNA.

In *chapter* 2 we analyze graph theory as basic tools to extract knowledge from networks. We present the graph theory in such a way we can use its definitions for successive needs.

The *chapter* 3 starts with the formalization of vector space on a graph in order to introduce in a more complete way the spectral graph theory. Throughout this chapter we focus the attention on the most important quantities and respective properties useful in network analysis.

In *chapter* 4, we focus the attention on the so-called complex network theory and its analytical concepts. This part is divided in many subsections each of them dedicated to a specific index or concept. This chapter contains the fundamental issues that we need to characterize the topological property of the real networks.

Chapter 5 concerns the most important statistical models for topological and dynamical network analysis. We implement some of these models in the following but the purpose of this part is to provide a deep overview on that models. In particular, we will deal with: Erdos and Renyi Random Graphs (and advanced in random graph modeling) and Random Walks based on Markov chain graph specification, Small World Networks and Scale Free Networks according to the Albert and Barabasi interpretation. We discuss also the application to the SNA of these models. The Exponential Random Graph model will be treated in chapter 6.

Indeed, *chapter* 6, deals with SNA in an advanced way, especially considering the exponential random graph model and its specification as Markov graphs and p* family models. Here in general, we will analyze the original contributions of SNA to the analysis of networks and also the points in which we think the complex and spectral perspective are useful.

Finally, *chapter* 7 is devoted to illustrate the ”meeting point” between SNA, from on hand, and complex networks models and spectral graph theory, on the other hand, by proposing some possible application. In particular we present some specific social network problem solved by means of these advanced tools. Several indices and models discussed in the previous chapters are here treated from a practical point of view.

CHAPTER 1

Elements of Network Theory

1.1 A brief history of network theory and network science

The last decades has witnessed an increasing interest about network structures in several scientific fields. This interest arises, first of all, because of the extremely *immanent* presence of networks in the real world. Indeed, network structures are everywhere: in physics to model discrete phenomena, in chemistry to represent the molecular structures, in engineering to illustrate the functioning of electric circuits, in information technology to study the world wide web topology and especially in social and behavioural sciences.

Networks are widely studied principally because they represent a powerful structural model for complex systems behaviour. In particular, a complex system is cosidered as an interacting ensemble of a multitude of elements whose interactions determine the macro-behaviour of the whole structure. Though the global system properties are generated by elementary interactions, its global features are definetly more complex than the summation of the single elements relational activities. The centrality of the network paradigm in physics derives from the fundamental role played by the study of elementary interactions among system components in system theory.

Though this importance, the earliest known study on network, has been developed in the field of pure mathematics. We are talking about the famous *Seven Bridges of*

Konigsberg written by Leonhard Euler in 1736 [21]. In Euler's problem the definition of the connections between the two Koninsberg islands and the mainland - vertices - by means of seven bridges - edges - is the first form of a graph. Moreover, the Euler's problem resolution[1], is the first appearance of some of the graph theory basic concepts as: degree and path (precisely, eulerian path).

In particular, while solving the Konigsberg problem, Euler took two important steps:

- The map of the city was replaced to a simple diagram showing its main features.
- He denoted four land areas as points A,B,C and D and seven bridges by lines a, b, c, d and e joining the points (see fig. 1.1).

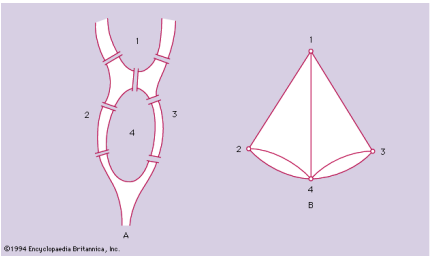


Figure 1.1: Seven bridges across the River Pregel flowing through the city of Konigsberg. The citizens wanted to find a route to visit all parts of the city by crossing all the bridges only once. This is the famous Konigsberg problem (source: Encyclopaedia Britannica inc. 1994)

Though the Euler's resolution method represents the first appearance of a systematic network analysis, in a description of the history of network theory and network science, it is better to keep separate the development of graph theory and its application (as the social network analysis or the electrical network theory). Indeed, the applications of graph theory to several scientific disciplines have represented the real motivation toward a development of a science of networks. Since its birth, graph theory rapidly became a branch of discrete mathematics and combinatorics devoted to the studies of the relations between elements located in a network structure. Graph theory has witnessed several exciting developments and has provided answers to a series of practical questions

[1] In the Euler's paper, the problem was based on the configuration of the city of Konigsberg. In particular, it concerns the possibility to find a route to visit all parts of the city by crossing the seven city's bridges only once.

such as: what is the maximum flow per unit time from source to sink in a network of pipes, how to color the regions of a map using the minimum number of colors so that neighboring regions receive different colors, or how to fill *n* jobs by *n* people with maximum total utility [12].

Approximately a century after Euler's work, graph theory began to be increasingly applied to various disciplines involving networks. The first application of graph theory outside of combinatorics was in chemistry [88]. In particular, the chemist Sylvester introduced the term "graph", referring to diagrams showing analogies between the chemical bonds in molecules and graphical representations of their mathematical invariants (see fig.1.2).

Another important application of graphs, probably more crucial for the development of network science, was the social interaction study of the early 1930's social scientists, in particular the studies of the Gestalt psychologist Jacob Moreno. During this period, Moreno creates the *sociogram*, a graph representation of social interactions among a set of actors (see fig. 6.1 on page 99). Though it was not explicitly formalized in terms of graph theory, this approach is of primary historical relevance for network theory application for at least two reasons. From one hand, because after his efforts also social scientists appreciate the importance of the patterns of connection among people for the study of the human society [**?**]; on the other hand, because Moreno's sociogram was the first application of network diagrams outside natural sciences and pure mathematics.

This attempt can be reasonably considered the "indirect" starting point of the increasing interest in network analysis, during the XX century. More directly Moreno's work represents the seed from which spread the new discipline, known as Social Network Analysis (SNA, from now on). SNA, in particular can be defined as the study of the social interaction through the *explicit* use of the graph theory tools and related statistical network models. Thus, it can be considered the most prolific meeting point between discrete mathematics, statistics and sociometry.

Since 1930's to nowadays, graph theory has been applied in various disciplines. Indeed, after the chemical graph structures and the social interactions modelling, in a number of other scientific fields the systematic implementation of the network models

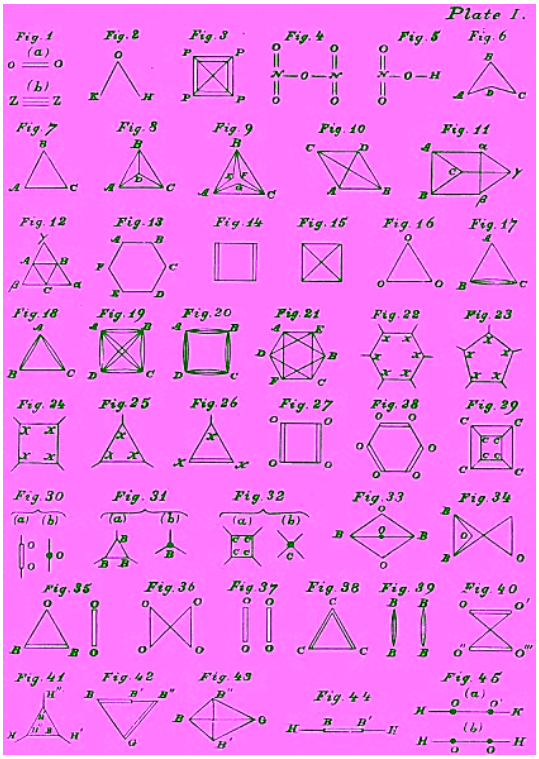


Figure 1.2: Reproductions of chemical graphs studied by Sylvester. It has been the first application of graphs outside the combinatorial analysis (source: Sylvester, 1878).

and graph theory starts. In general, one can refer to these graph-based disciplines with the general term of *network analysis*. Indeed, several common features have been discovered among the networks studied in sociology, psychology, physics, and so on (see for an introduction to these interdisciplinary connections [6] or [**?**]). The common features in the network modelling, across the singles disciplines, is one of the main motivations of this dissertation.

More recently the peak in the ”network interest” has been strongly influenced by the fast development of the World Wide Web, perhaps the most dynamic and pervasive ”network” in the recent human history. The Internet can be viewed as an huge networks of connected computers, connected web pages and/or connected users and hubs. Viewed in this form the web implies: machines, documents and human complex interactions. For this reason the Web, in network analysis, has a double role: it can be considered as object of study (in particular in the network information technology and engineering) and, on the other hand, it can be considered a sort of archetypal structure to model other real networks (see in particular the works on the power law distributions of Albert Barabasi in [6]).

The increasing implementations of network perspective in different disciplines, the birth of concrete networks itself and the discrete modelisation for the behaviour of connected elements in complex systems, led to the development of a total new discipline called: *complex networks theory* (sometimes denoted as CNT). Complex networks theory arises especially in physics, though it can be considered a completely general characterisation of any kind of network having particular topological and dynamical features. The use of the term ”complex networks” depends on the fact that the networks studied in CNT contain much randomness, but to model them as purely random graphs (cfr. section 5.1) would not be very apt since they also contain structure. During the last decades CNT has been applied to a multitude of real cases with very nice results. Indeed, the success of complex network theory, as new way of study complex phenomena, is due to the discovery that this approach overpassed the range of the ”‘simplest” systems and became a very prolific way of studying huge networks and even social interactions and biological processes.

This multidisciplinarity, and the popularity of the study of ”connections”, led network analysis toward a sort of unification under a unique new science: the so-called *network science*. It is possible to succintly define network science as a new scientific field studying the interconnections among different physical, informational, biological, cognitive, and social networks under a general point of view. One of the central themes in network science is the extraction of information hidden within the wiring of a network [49]. Briefly, network science seeks to discover common principles, algorithms and tools that govern network behavior, independently from the field in which the network is observed.

The main research questions in network science can be summarized in three points:

- What are the relevant structures in the networks?
- What processes give rise to these structures?
- How are dynamical processes affected by the network structure?

The first question concerns the study of the structure of a network through their structural indices, many of them developed by statisticians, that allow us to describe and compare networks in several ways. The other questions deal with the networks dynamical behaviour, their growth (i.e. the modifications of links, the probabilities of increasing/decreasing of connections given some actor's characteristic, etc.) and the processes naturally born in a relational context (i.e. the disease spreading, the game theoretical models of social stability and other social interactions phenomena).

1.2 Networks and Complexity

In the current language the concept of *complexity* has been related to several notions. Perhaps the earliest (and simplest) but most used notion of complexity is simply that of the cardinality and/or differentiation of an object's components. Organisms, organizations, societies, or technical systems with many parts are more ”complex” than those with fewer components. Such a straightforward conception of complexity has been widely employed in organization theory and biology, and widely appears in some of the earliest of sociological works (e.g. the works of Durkheim especially [29]).

Complexity as cardinality is strongly related to another concept which is closer to the study of dynamics in general systems. In particular, it is common in this field to speak of systems of high dimensionality as being complex. Examinations of many families of such systems have revealed often surprising behaviors, such as unexpected convergence to stable fixed points of a variety composed of elementary parts.

The concept of ”parts”, ”interactions” and collective behaviour are the basic concepts on which the complex systems and the networks meet each other. Some author affirms that every complex system can be reduced to a network and can be usefully studied by modelling systems as networks. This concept starts from considering that systems in nature are discrete structures even if they look like continuous ensemble. The quantum mechanics is perhaps the most important example of such way of interpret the reality.

The importance of specification of networks as complex system allows us to claim the application of different tools to the study of social network interactions. Indeed it is possible to define a complex social systems. We intend with this concept the networks of collective action which manifest a degree of internal coordination and integration such that the behaviour of their individual members appears to be connected in some way [25]. It should be noted that these systems may be non-linear and non-equilibrium, i.e. interaction effects can be multi-directional and move across different levels of organization, and that the enesemble they form may be in a constant state of change, never reaching a settled (equilibrium) state [20]. The most important aspect for our purposes is that the networks involved in these systems are naturally occurring and socially consequential, not artificially generated for experimental manipulations. Then we will use the above definition to deal with social dynamics in a network complex ssystem form and consequently to applying

1.3 Complex networks, graph theory and social network analysis: a new perspective

Let consider in this section the specific characteristics of the complex network theory framework. It has been already mentioned that one of the most important impulse for network theories unification, has been the definition of *complex networks* and the applicability of the network persepective to a multitude of real-world strucutures. This scientific characterization, for the first time since the earlier graph theory works, considers the network as a general representation of a system in which elementary units interact by means of links among them. This representation allows to model: social actors and their relationships, web pages and their links, neurons and their synapses, etc.

A noteworthy number of researchers affirm that complex networks is not only a widely used concepts but actually it is a field where dramatic advances have been witnessed in the past few years. The core of the analysis is the structural and dynamical properties of networks on a general level, rather independently from the explicit nature of the units (nodes) and their linkages.

Though network science is a concrete field of research as we discussed previously, we have to admit that some additional effort must be directed toward a real convergence of the network analyses. For example, in SNA the ”social analysts” focus most of the times on the local configurations within a social network, whereas in physics the analysis of the same network the attention is toward the globality of the system rather than in regional aspects. There is not an analytical reason for these differences. It is just a matter of different adopted paradigms. Indeed, SNA is influenced by the sociological theoretical framework, principally derived from the works on social interactions of Georg Simmel [79]: in this scheme, is prevalent the structuralism and the whole social groups is mainly formed by the social interactions that are observed in dyads and triads. Physical network study is instead oriented toward the analysis of the whole object (namely, the system) that is only theoretically decomposable in a number of interacting elements. In particular, in according to the physics viewpoint, the local dynamics cannot

describe the whole system functioning because its global behaviour is more complex then the summation of the parts.

One of the topic of this dissertation is to give a contribution to reduce the distance between these approaches. In particular, it is not a reductionism to consider a social network as a system whose actions are not determined by a simple summation of their elements. For example an individuals have the probability of activate some relationship with someone else in the network independently on the number of dyads and triads in which it is involved. The probability of connections depend on other features of the network rather than on the local configurations. However some actor's characteristics can be usefully interpret in terms of local connectivity: for example the prestige of a subject in a group can succesfully measured by the kind of dyads and triads he participate.

In practice we would like to emphasize that a network study on the same object can be conducted by using both these approaches. But we think that could be even better consider them as complementary viewpoints and impement a mix of both in network analysis.

Then, the lack of interconnections between complex network studies and network theory, form one side, and social network analysis, on the other side does not lie in the different characterstics of the networks under study but it arises from the different paradigms used in the reality interpretation.

Throughout this dissertation we will start from the definition of *complex network* in order to identify avery object that can be represented by interconnected elements. Indeed, a complex network is an object that can be seen as a graph with non trivial topology and that can be fully described by means of its structural characteristics and/or by the analysis of its dynamical behaviour. Throughout this work the distinction will be made if it is necessary. Otherwise both statics and evolution of complex networks are considered interrelated elements that jointly describe the system's behaviour and the topological features.

CHAPTER 2

Basic and advanced notions on graph theory

The purpose of this chapter is to furnish the fundamental on graph theory, which represents the mathematical foundation of network analysis and the necessary prerequisite to understand advanced theory. The topics covered in this chapter are indeed important for the methods discussed in the following.

Graph theory is the natural theoretical framework for the mathematical treatment of complex network. As we states previously, graph theory is the basic mathematical background necessary for network science applications Therefore, graph theory as the set of the theories and concepts on which complex network theory has been developed. Formally, a complex network can be represented as a graph: we will indicate a (complex) *network* also as a *graph*.

In this chapter we will describe all the kinds of object studied by graph theorists, illustrating their most basic and useful characteristics, regardless their indices and related models (which are treated in detail in the chapters 4 and 6). In particular, in the first part of this chapter we will introduce the general graph theory framework, that is common to the various types of network analyses. We then introduce the specific definitions that are necessary to approach in a formal fashion the network analyses.

2.1 General concepts in graph theory

A *graph* is a mathematical entity $G = (V, E)$ composed of two sets: a set $V \equiv \{v_1, v_2, ..., v_N\}$, of cardinality $|V| = N$, containing *nodes* (also known as *vertices* or *points* or, especially in SNA, *actors*) and a set $E \equiv \{e_1, e_2, ..., e_K\}$, of cardinality $|E| = K$, containing *edges* (or *links* or *lines* or *arcs*). In particular, the cardinality of the node-set of a graph *G* is called the *order* of *G* and obviously must be $N \neq 0$ or, in other terms $V \neq \emptyset$. Instead, cardinality of the edge-set *E* (i.e. the number of edges) is called the *size K* of *G* and it variates between 0 and $\binom{N}{2}$. If $E = \emptyset$ then $K = 0$ indicating a totally *disconnected graph*; in the case of *K* equalize its theoretical maximum value we identify a *complete connected* graph. By comparing the cardinality of both *V* and *E* (or equivalently the order and the size), a graph is said to be *sparse* if $K << N^2$ and *dense* if $K = O(N^2)$.

Specifically, a graph of order $N = n$ and a size equal to $\binom{n}{2}$ is called *complete n-graph* and denoted as K_n in which every node is connected with each other. In order to build up an algebra of graphs (see section 3.1) and to starts an induction, it is possible to introduce two special objects: the *empty graph* or *null graph* with no nodes and hence no edges denoted as $\emptyset$ and the *trivial graph* (a graph containing exactly one node) denoted as K_1. In particular, we can also have a family of empty null edgeset graphs with *n* disconnected nodes, of which the trivial graph is the simplest element.

Basically we need to distinguish between three kinds of graphs depending on the property of the edge-set: i) if the elements of the set *E* consist in unordered pairs of nodes we will indicate this object as *undirected graph*; ii) if *E* contains ordered couples of nodes, we will speak of *directed graphs* (or *digraphs*); iii) the generalization of these previous objects is called *weighted graph* and it is obtained when for each element e_{ij} of *E* there exists a one-to-one correspondence with the element w_{ij} of another set *W* (the weights-set), composed of real numbers called *edge weights* (or simply *weights*). In all the three cases, if a pair of nodes $v_i, v_j \in V$ is connected by an edge, the element $e_h = \{v_i, v_j\}$ belongs to the set *E*, where *h = i,j*[1]. Unless otherwise stated we will deal

[1] In describing a general graph *G*, we will use most of the times the double index notation e_{ij} (where for undirected graphs hold the relation $e_{ij} = e_{ji}$); while the multi-index notation e_h where $h = i, j$ if *G*

only with labelled graphs on $N = n$ nodes. Sometimes, for convenience we will call the nodes in *V* by using their labels *i* with $i = 1, ..., n$. Therefore, for undirected graphs, the edges (couples of nodes) will be simply indicate as $\{i, j\}$ while for digraph we will denote it as (i, j) [2]. Moreover, if it is necessary to specify the cardinality of *V* and *E* a graph will also be noted by $G(N, K)$ or $G_{N,K}$ and, in non ambiguous context, simply with *G*.

An edge connecting two nodes *i* and *j* is said to be *incident* with that nodes. Two nodes joined by an edge are referred to as *adjacent* or *neighboring*. A node incident with no edge is called *isolated node*.

There exist several ways to represent a graph. The natural way is to specify the elements of the two sets by labelling the nodes in *V*. In other words, in this way a graph is given by its node-list and edge-list. However, the more parsimonious way to represent a graph is drawing it by a diagram in which nodes are represented by dots and edges are depicted by a line joining two dots if and only if the two corresponding nodes are connected (i.e. if the couple of nodes belongs to the set *E*). For digraphs a directed edge is depicted as an arrow; while for weighted graphs to each edge it is attached the corresponding weight in the set *W*. Perhaps the most useful graph representation is by using some graph associated matrix. We will go back later on this point in section 2.4.

In order to illustrate an usual graph representation (undirected in this case) let consider the graph $G(V, E) = G(9, 18)$ in fig. 2.1 in which the node-set is of cardinality $N = 9$ (the order of *G*) where the nodes are labelled by numbers form 1 to 9. *G* has the following edge-set:

$E = \{e_{12}, e_{23}, e_{34}, e_{45}, e_{56}, e_{61}, e_{17}, e_{72}, e_{29}, e_{95}, e_{57}, e_{74}, e_{84}, e_{83}, e_{39}, e_{96}, e_{68}, e_{81}\}$.

Given that the cardinality of *E* is equal to $|E| = 18$, the size of *G* is $K = 18$. In the figure 2.1 we represent the graph by means of a diagram.

In general, in drawing a graph, the position of dots and lines is irrelevant because the crucial element is the connection between the elements of *V* and nothing else. More

is undirected or $h = (i, j)$ is directed (see section 2.2) will be used in all the cases in which we would like to simplify the notations.

[2] We will use the algebraic notation to distinguish between ordered and unordered pairs of elements of a set: in particular, if a couple of nodes *i* and *j* is in curly brackets it indicates an unordered relation; while if the pair is in round brackets it denotes an ordered relation.

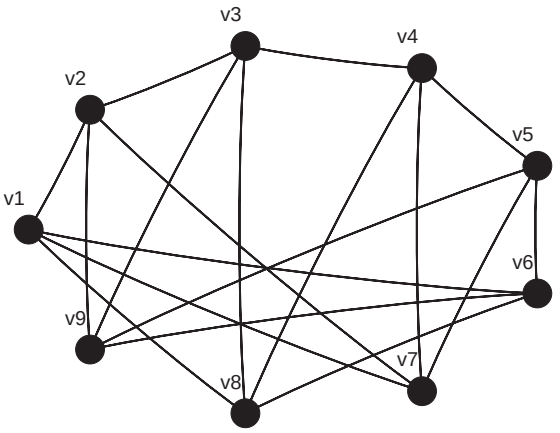


Figure 2.1: An undirected regular graph G with 9 nodes.

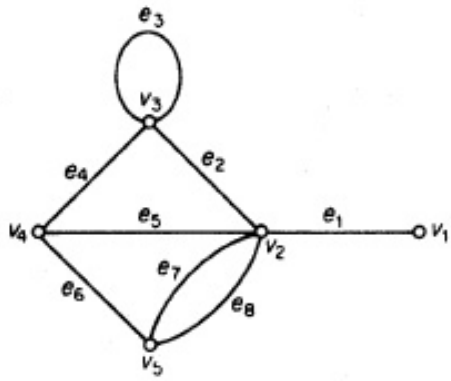


Figure 2.2: *A multigraph with one loop and a multiple edges. Here we labelled also the egese to emphasized the multiplicity of the relations (Source: Bondy and Murty, 1976).*

precisely, *graph topology* is only characterized by its node connectivity.

The graphs described until now are canonical. Beside these types, there exist also special kinds of graphs: in particular, it is possible to define graphs containing *loops* (i.e. edge having as starting and ending point the same node) and *multiple edges* (i.e. couple of nodes connected by more than one edge). Graphs presenting either of these features are called *multigraphs* (see fig. 2.2). For an introduction on multigraphs see [46]. Generally in this work, unless otherwise stated, we will deal with canonical graph.

An important concept is the one of *graph isomorphism*: two graphs $G(V, E)$ and $G_1(V_1, E_1)$ are said to be isomorphic if there is a biiection $\phi : V \rightarrow V_1$ such that $e_{ij} \in E$ if and only if $\phi(e_{ij}) \in E_1$. The relation ϕ preserves the adjacency (see fig. 2.3). Isomorphic graphs have the same size and the same order. Usually isomorphic graphs are indistinguishable, unless they have labelled set of vertices. Therefore, if *G* and *H* are isomorphic graphs we will write: $G \cong H$ or simply $G = H$. An isomorphism from a graph to itself is called an *automorphism*. An automorphism is a *permutation* of the vertices of *V* that maps edges to edges and nonedges to nonedges [44], i.e. it preserves incindences between nodes and edges (for more detail on graph automorphism see [92]). The concept of automorphism is fundamental in many applications, e.g. in the ERGM parameters estimation (see chapter 6). In particular, a set of automorphisms is

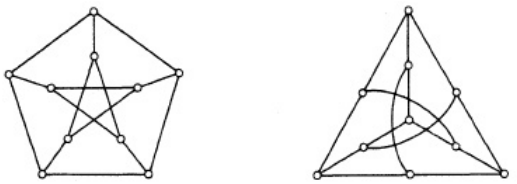


Figure 2.3: The graphs G(V,E) and H(V*,E*) are isomorphic because the adjacency of the nodes in V and V* is preserved (Source: Bondy and Murty, 1976)

a group (called *automorphism group* of *G*, when it is equipped with the operation of composition of functions) and it allows us to generate a family of graphs with the same structure by mean of special invariant quantities (the node degrees of a graph are the most important examples of such invariants, cfr. in section 4.1.2 lemma 1 on page 56).

2.1.1 Subgraphs

A *subgraph* $G' = (V', E')$ of $G = (V, E)$ is a graph generated from *G* and such that $V' \subseteq V$ and $E' \subseteq E$. Any generic subgraph may not include all edges between the nodes in the subgraph. However there are some peculiar kinds of subgraph. In particular, if *G'* contains all the edges that join any pairs of nodes in *V'*, then *G'* is called *subgraph induced* by *V'* and is denoted by $G' = G[V']$. Such a subgraph is induced by *V'*, since the subset of nodes has generated the subgraph. Moreover, if *G'* contains all the nodes incident with the edges in *E'*, then this subgraph is called induced by *E'*, because is edge-generated. It is denoted by $G' = G[E']$. Note that an edge-induced subgraph of *G* does not admit isolated nodes. Given an edge-induced subgraph $G' = (V', E')$ to respect a graph $G(V, E)$, the subgraph $G'' = (V, E - E')$ is called the (*edge-*)*complement* of G' respect to *G* [**?**]. If *V'*=*V* then we call it a *spanning* subgraph of *G* (see fig. 2.4. Note that a node-induced subgraph may have isolated nodes.

A special subgraph, often used in complex network and social network analysis, is the subgraph of the neighbours of a certain node *i*, indexed as G_i or $G(i)$ and defined as the subgraph induced by V_i which is the set of all the nodes adjacent with *i*, i.e. the graph

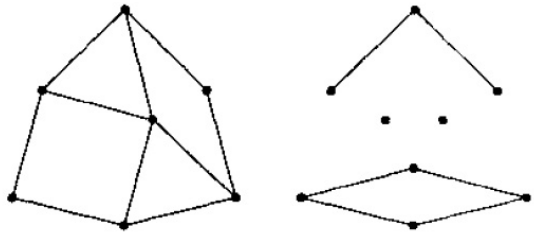


Figure 2.4: *From left to right: a subgraph of the graph G in fig. 2.1 induced by the nodes 1,2,3,4,5,6,9 and a spanned subgraph of G (Source: Bollobas, 1998)*

$G_i = G[V_i]$. Again, very interesting subgraphs are the *dyads* and the *triads*: the former represents a subgraph induced by only a pair of nodes of the node-set *V* of a graph; the latter is another node-induced subgraph formed by three nodes of the node-set *V*. Since, both dyads and triads are node-induced subgraphs, they are defined as a subset of *V* and all edges between pairs of nodes in the subset.

A subgraph is said to be *maximal* with respect to a given property if it cannot be extended without losing that property. This latter property is inherent some structural or dynamical characteristic of the network.

2.1.2 Walks, Trails and Paths

In a graph nodes can be basically connected pairwise, this is the concept of adjacency. However, there are other ways in which two nodes can be connected by ”indirected” links that pass through other nodes in the graph. These properties are the basic ones for many other more complex properties. In particular, *walks* and *paths* will allows us to compute the distance between pairs of nodes and they are the basis of the concepts of geodesic, diameter and eccentricity. A *walk* W is a sequence of nodes and edges, in which each node is incident with the edges that follow and precede it in the sequence. The length of a walk is the number of edges in the sequence. Clearly, if an edge is repeated more then once it is counted each time it occurs. The *inverse of a walk*, denoted with W^{-1} is simply the walk W listed in the opposite order. In unordered graph, given

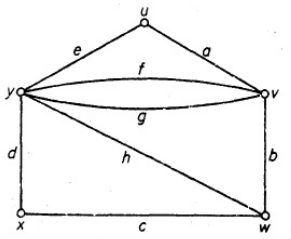


Figure 1.8

Figure 2.5: *Examples of a walk, a trail and a path on a given graph (source: Bondy and Murty, 1976)*

the non-ambiguity of the connection between each pair of nodes, one can list only the nodes involved in the walk, ignoring the edges. More precisely, a walk of length *n* ($n > 0$) is a sequence of *n* edges and *n*+1 vertices: $W \equiv \{v_0; (e_{01}); v_1; (e_{12}); ...; (e_n); v_n\}$ (from v_0 to v_n). The walk connects v_0 and v_n. In the following we will give the most important specification of a walk:

- A walk is a *trail T* if no edge occurs more than once (though vertices may be repeated).
- A trail/walk where $v_0 = v_n$ is *closed*.
- A *path P* is a trail in which no vertex is repeated except possibly the initial and end points.
- A closed path, of at least three nodes, is called a *cycle C*.

In particular, cycles are important in the study of many network properties, as in measurement of balance and clusterability. Special case of cycles are the cycle C_3, the *triangle*, formed by only three nodes, and the *Hamiltonian cycle*, when every node in the graph is included exactly once in *C*. Specifically, the property of containing at least a triangle is a quite common characteristics of several graphs: indeed, the well known Mantel's theorem (see [54] for a detailed proof and the its consequences in extremal graph theory) affirms that every graph of order $N > 2$ and of size equal to $K = N^2/4$ contains a triangle. This is a useful result for the study on the triads.
Other special closed walks are those that include every edge [57]: in this category we

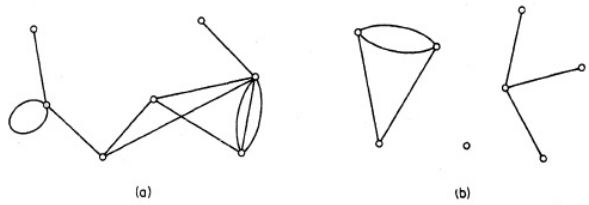


Figure 2.6: *(a) A connected graph (one component); (b) A disconnected graph with three components (source: Bondy and Murty, 1976)*

can include the so-called *eulerian trails* (which are those trails in which every edge is contained exactly once).

2.1.3 Connected graphs and components

An important property of a graph is wheter or not it is connected. A graph $G(V,E)$ is *connected* if there is a path between every pair of nodes in V, i.e. every node is reachable from an arbitrary one (see fig.2.6). The notion of connectedness is fundamental in several network model, like the ones based on Markov Chain and Random Walk. A graph is *disconnected* if it is not connected. A disconnected graph can be partitioned in *components*, that are blocks of connected subgraph of G: more precisely, a component is a maximal connected subgraph (see fig.2.6). Clearly a connected graph consists of one component. A totally disconnected graph is isomorphic to the trivial graph: it is simply the graph composed of only isolated nodes. The number of components of a totally disocnnected graph is equal to N the number of nodes in V.

2.1.4 Distance in a graph

Topology of a graph is completely defined by considering its connectivity, therefore also the notion of distance (metric) on a graph is related to the connections between nodes. As mentioned before, the connection between two nodes i and j is defined by the path connecting them. The length of the path is simply the number of edges in the sequence. However, in general there exist several paths of different length between two

arbitrary nodes *i* and *j*. The *shortest path* between *i* and *j* is called *geodesic distance* and it represents the "natural" metric in a graph[3]. We will denote the geodesic between *i* and *j* as $g(i,j)$ or g_{ij}. If there is no path between two nodes (i.e. they belong to different graph components) the geodesic distance is infinite (or undefined). Hence, if and only if *G* is an undirected and connected graph, then the set *V* of vertices of *G*, equipped with the geodesic distance $g(i,j)$, represents a *metric space*.

The *eccentricity* (or *association number*) of a node is the largest geodesic distance between that node and any other node: formally, $ecc_i = max_j g(i,j)$. The minimum value of eccentricity of a node *i* is 1 (when *i* is adjacent to all the other nodes in *V*); the maximum value can be equal to *N*-1. Several measures of centrality, such as the center and the centroid of a graph, are based on the eccentricity of a node [91].

Considering all the nodes in *V* of a connected graph, the largest eccentricity measured among them is called the *diameter* of the graph. Equivalently the diameter of a graph is the largest geodesic measured between any pair of nodes. Formally: $Diam(G) = max_i max_j g(i,j)$. If a graph is not connected the diameter is infinite. This index quantify how far apart the farthest couple of nodes are. The diameter of a graph is closely related, as several other graph parameters, to the eigenvalues of the laplacian matrix of G (cfr. sect 3.2.1).

2.1.5 Connectivity of a graph

The connectivity of a graph is function of the attitude of a graph to remain connected when nodes and/or edges are removed from their respectively sets. A node v_i of a graph *G* is a *cutpoint* if its removal disconnects the graph: i.e. the number of components of *G* increases at least of one. A cutpoint of *G* is a subset *V'* of *V* such that $G(V - V', E)$ is disconnected. A *k*-cutpoint is a cutpoint of *k* elements of *V*. This concept is strongly related to the one in general topology, in which a cutpoint is a point (or a subset) of a connected topological space such that its removal causes the disconnection of that space. Similarly, an edge is a *bridge* if the removal of the edge disconnects the graph.

[3]Geodesic is the induced distance on discrete surfaces. However, it is not a real-valued distance because its codomain is the set of the natural numbers

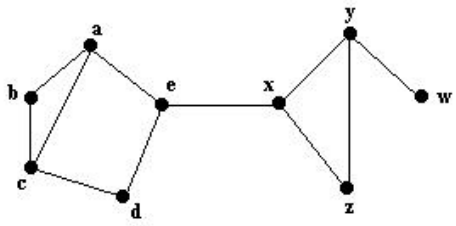


Figure 2.7: A graph in which a possible cutpoints are the nodes e and x; a bridge is the edge (ex)

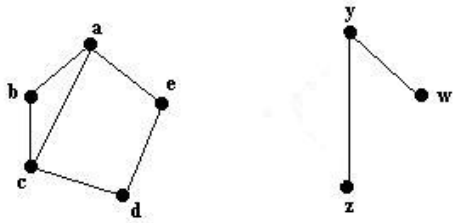


Figure 2.8: The effect of the removal of the node x is to increase the number of components of the graph.

An *l*-edge cut is an edge cut of *l* elements. A bridge is a 1-edge cut (see sect. **??**). If G is nontrivial and *E'* is an edge cut of *G*, then $G(V, E - E')$ is disconnected.

Let consider the graph in fig. 2.7.

A possible cutpoints are the nodes e and x, because their removal increases the components of the graph of one. In figure 2.8 we see the effect of the removal of the node x: now the graph G has two components and it is disconnected (the same if we operate the removal of the node e).

The only bridge in the graph in fig. 2.7 is the edge e_{ex} connecting the nodes e and x (see fig. 2.9).

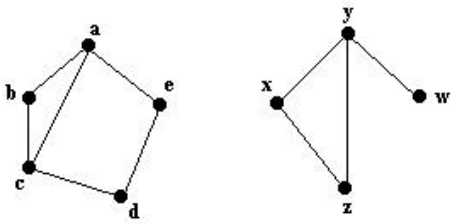


Figure 2.9: The effect of the removal of the bridge (ex) is to increase the number of components of the graph.

Node-connectivity and edge-connectivity are quantities based on the notions of cutpoints and bridges. They are crucial in partitioning the graph and in measuring the cohesiveness of the network. The *node-connectivity* of a graph *G*, denoted as $\kappa(G)$ is the minimum number of nodes that must be removed to disconnect the graph. For example, if a graph contains a cutpoint then $\kappa = 1$; if we need to remove a couple of nodes to make the graph diconnected then $\kappa = 2$. In particular, $\kappa(G) = 0$ if *G* is either trivial or disconnected. *G* is said to be *k*-connected if $\kappa(G) \geq k$. Therefore, removing any number of nodes less than κ does not make the graph disconnected.

Edge-connectivity of a graph $\kappa'(G)$ is the minimum *l* for which *G* has a *l*-edge cut. If *G* is trivial, $\kappa'(G)$ is defined to be zero. If $\kappa'(G) \geq l$, *G* is said to be *l*-edge connected.

2.1.6 Cuts and cutsets

A generalization of the discussion in section 2.1.5 consists in considering, for a connected graph $G(V, E)$, a generic *partition* of the nodeset *V* into two disjoint subsets V_1 and V_2 such that $V = V_1 \cup V_2$. In particular, the set of all those edges of *G* having an ending node in V_1 and the other in V_2 is called a *cut* and it is denoted as $S = \langle V_1, V_2 \rangle$. Similarly to the bridges case, the removal of the edges in a cut of a connected graph *G* will disconnect the graph. A cut is called a *cutset* if the removal of the edges in the

cut results in a disconnected graph with *exactly two components*. Specifically, we can consider the cutset as a collection of bridges whose removal disconnect the graph into the components V_1 and V_2. Equivalently, the cutset is the minimal collection of edges whose removal disconnect the graph.

2.2 Directed graphs

Directed graph or, shortly *Digraph*, distinguish itself by the simple graph only for the elements in *E*. Indeed, in these kinds of graph the order of the connection between two nodes is important, in particular: $e_{ij} \neq e_{ji}$. Hence a digraph is an object $G_d = (V, E_d)$ in which the set E_d is the set of *ordered pairs* of elements of *V*. If there is no ambiguity in the notation we will denote also the digraph simply with *G*. In particular, in a digraph edge $e_{ij} = (i, j)$ [4], we have to distinguish the role of the two nodes: *i* is referred to as *starting node* (or sender) and *j* as *ending node* (or receiver). Since this difference between the two nodes in the pairs, the concept of adjacency in digraph is somewhat complicated than in the simple graphs [91]. When a digraph is represented with a diagram, the edges e_{ij} are depicted as arrows starting form *i* and pointing to *j*.

In practice, all concepts defined for graphs and subgraphs can be extended in the case of digraphs: the only additional element to consider is the direction of the relations. For example in the case of walks, paths and trail the orientation of connections must be take into account when we apply those concepts to the directed graphs.

Considering the special subgraphs defined before, namely dyads and triads, the direction of the edges influences their possible states. In particular, whereas the dyads in undirected graphs have only two possible state (absence or presence of the edge), in digraphs the dyads could assume in general four possible states (called isomorphism classes). The number of isomorphism classes of triads increases to 16. We will go back to these concepts later, in the chapter dedicated to the social networks.

[4] Note that in this case we use the round brackets in place of the curly brackets, to indicate the importance of the ordering in the couple *i*,*j*

2.2.1 Directed Walks, Paths, Semipaths

A *directed walk* is a sequence of nodes and edges such that each edge has its starting point at the previous node and its ending point at the following node. The length is measured in the same way that for the simple graphs: it is the number of occurence of the edges in it. A *directed trail* is a walk in which no arcs is included more than once. Finally, the *directed path* consists in a directed walk where no node and no arc is included more than once. If, in a directed path, the initial and the terminal node coincides we will call it a *closed directed path*. Thus in the directed versions of these concepts the directed walks/trails/paths consist in edges that point in the same direction. A *cycle* in a digraph is every closed directed path in which there are at least three nodes. If one removes the restriction of the direction in a walk it is possible to define a so-called *semiwalk* in which the edges in the sequence may point in either direction [91]. A *semipath* is a sequence of nodes in which no node is repeated more than once and where the direcion of the edges is irrelevant. A *semicycle* in a digraph is a closed semipath composed of at least three nodes.

2.2.2 Connectivity in digraphs

The emphasis on the direction modifies the problems of reachbilty and connectivity. In particular, given two vertices v_i and v_j one can distinguish between:

1. *weakly connected* nodes if they are joined by a semipath
2. *unilaterally connected* if they are connected by a directed path from v_i to v_j or a directed path from v_j to v_i
3. *strongly connected* if they are connected by a directed path from v_i to v_j and a directed path from v_j to v_i
4. *recursively connected* if they are strongly connected and the directed path from v_i to v_j includes exactly the same node of the directed path from v_j to v_i

Consequently, a digraph *G* with node-set *V* will be indicated as weakly, unilaterally, strongly or recursively connected if all pairs of nodes in *V* are weakly, unilaterally, strongly or recursively connected.

The notion of distance in a digraph is also measured in terms of geodesic distance: the only difference, in comparison to the undirected case, is to consider only the edge that point in the same direction. Thus, in digraph the geodesic between two nodes v_j and v_i can be different if we consider the path from v_j to v_i or its reverse from v_i to v_j. Therefore we can state that $g(i,j) \neq g(j,i)$. This fact implies that in digraph the geodesic is not a distance but a *quasimetric*[5].

2.2.3 Generalized digraphs: weighted graphs

Weighted graph is the most general type of graph and it is defined as the graph $G = (V, E, W)$, where the set *W* contains the *values* (*weights*): $W = \{w_1, ..., w_K\}$ associated with each edge in *E*. Where: $w_i \in \Re$. In order to avoid ambiguity, a weighted graph is sometimes indicated as G^W.

The edge associated value is variously called as its cost, weight, length or other term depending on the application. Such graphs arise naturally in many contexts, for example in optimal routing problems such as the traveling salesman problem.

In particular in weghted graph the length of a path *P* starting from the node v_0 and ending at the node v_{k-1} is the sum of the weighted edges w_i contained in the sequence:

$$w(P) = \sum_{i=0}^{k-1} w_i \tag{2.1}$$

The distance between two nodes v_i and v_j, is equal to the minimum weighted length connecting the two nodes. Indeed, because we have to take into account the edge values, in the case of the weighted graphs the distance is, in general, a real valued function. Therefore is not a metric because the orientation of the edges may causes the relaxation of the property of symmetry[6]. Another reason for which geodesic in valued graphs is

[5] The quasimetric is a metric derived from the relaxation of the axiom of the symmetry.

[6] In general also for valued graphs in which the set *W* is a subset of the real numbers, we have the

not a metric depends on the fact that the distance between two nodes can be undefined even if there exists a path between them. Indeed, it could happen that the sums of the values associated to the edges in the path may sum up to a negative value. Negative values implies the relaxation of the first axiom for the metric (the non-negativity), hence it is not a distance[7]. One special application of weighted graphs is the set of graphs whose values consists of probabilities. These graphs are known as Markov chains graphs, and they are related to their transition matrices in the same way simple graphs are related to their adjacency matrices (see section 2.4). In these graphs the sum of the values for all the edges incident with the nodes is constrained to be equal to 1. Such structures are very important in several stochastic models for networks, therefore we will back on Markov chains in the following.

2.3 Special graphs: Bipartite graphs and trees

2.3.1 Bipartite graphs

A *bipartite graph* is an object $G(V_1, V_2, E)$ composed of two disjoint sets of nodes. Thus is valid: $V_1 \cap V_2 = \emptyset$. Theorem 1 shows a necessary and sufficient condition in order to identifies bipartite graphs by the analysis of the cycles in it [15] [8].

THEOREM 1: *A graph is said to be bipartite if and only if does not contain an odd cycle.*

Proof. supposing *G* a bipartite graph with nodesets V_1 and V_2 and let $v_1, v_2, ..., v_s$ be a cycle in *G*; assuming $v_1 \in V_1$, hence $v_2 \in V_2$, $v_3 \in V_3$, then $v_i \in V_i$ if and only if *i* is odd. Suppose now that *G* does not contain an odd cycle. Let choose a node in $V = V_1 \cup V_2$, supposing *G* connected, and assume that $V_1 = \{v_i : d(i,j) \in O\}$, where

inequality: $g(i,j) \neq g(j,i)$

[7] Several restrictions can be placed on the values in the set *W*. For example, Harary proposed to consider as a network only the valued graphs where $w_i \in \Re^+$ [46]; Roberts, proposed only integer values allowed for the elements of *W* [71].

[8] This is the first example of the importance of cycles to identify uderlying structures of a graph. We will use more intesively the notion of cycles in the following.

O is the set of the odd natural numbers. We know that $V_2 = V - V_1$ since there is no edge connecting the nodes of the same nodeset V_k, with $k = 1, 2$, otherwise it would be admissible the existence of odd cycle in *G* (this would contradict the hypothesis). Hence *G* is bipartite [15]. □

Combining theorem 1 and the Mantel's theorem (see section 2.1.2) we can find a lower bound for the size of a graph *G*: under the size $\lfloor N^2/4 \rfloor$ (where the symbol $\lfloor x \rfloor$ is the integer part of the real number x) a graph *G* must be bipartite because it is the maximal size of a graph of order *N* containing no odd cycles [15]. Indeed the triangle is the simplest odd cycle that one can observe in given graph. A graph with size exactly equal to $\lfloor N^2/4 \rfloor$ is called *complete bipartite graph* because is the maximal connected bipartite graph.

Bipartite graphs represent a well known case study in social network analysis: in particular, they arise in the two-mode network data in which there are two sets of actors and a relation linking the actors in a set with the ones in the other set[9] [82].

For a given graph $G(V, E)$, the partitioning of the nodeset *V* can be generalized from a 2-partition (the case of bipartite graph) to an *s*-partition. In particular, if the nodeset *V* of *G* can be partitioned in *s* disjoint subsets, such that $V = V_1 \cup V_2 \cup \ldots \cup V_s$ where $V_1 \cap V_2 \cap \ldots \cap V_s = \emptyset$, we will define such graph an *s*-partite graph.

2.3.2 Trees and forests

We saw that cycles are very important connectivity issue in several graphs: for example the order of cycles allows us to recognize bipartition in the nodeset (several features of cycles will be clear in the following). However there exist kinds of graph (both simple or directed) in which no cycles are allowed: such peculiar graphs are known as *trees* and denoted as $t(V', E')$ in which V' and E' can often be considered as subsets of the nodeset and the edgeset of a graph $G\,(V, E)$ (see fig.2.10). In particular, given that trees are acyclical graphs, they contain the minimum number of edges necessary to the graph to be connected. For this reason every lines in a tree is a bridge.

[9] The classical example is the case of the marriage system in a community: the sets are respectively the set of husbands and the set of wives.

Sometimes it is convenient to consider one vertex of a tree as special node: such a vertex is called the *root* of the tree and it can be considered the starting node of *t*. In particular, if we deal with directed edges, the root is the only node in *t* having indegree equal to zero.

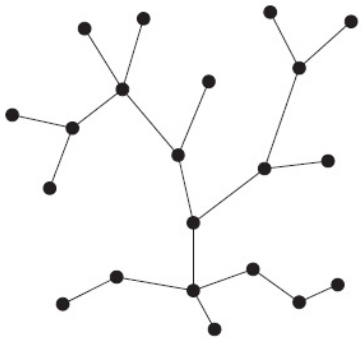


Figure 2.10: A tree.

An important notion is the one of *spanning tree T*: given a graph $G(V, E)$, a spanning tree is a tree containing as nodeset exactly *V*: $T(V', E')$, where $V' \equiv V$ and $E' \subset E$. In particular, the following proposition holds: every connected graph contains a spanning tree, with any specified vertex as its root (see [28] for a detailed proof). Briefly, for a given graph *G* with one component (a connected graph), a spanning tree is the minimal connected subgraph of maximum order. Such a structure arises naturally in many otimization problems and it is sometimes the basis of several network characteristics. The complement (see definition in 2.1.1) of a spanning tree *T* in *G* is called a *cospanning tree* $\bar{T}$ of *G*. The edges of a spanning tree *T* are called the *branches* of *T*, the edges of a cospanning tree $\bar{T}$ are called the *chords* of *T*. As consequence, if *G* is a graph with $K = N - 1$ edges, the following three statements are equivalent: (a) *G* is connected; (b) *G* is acyclic; (c) *G* is a tree.

A disconnected graph with *r* components and containing no cycles is called a *forest*, that is a graph composed of *r* trees, with $r \geq 2$ (see fig.2.11). A *spanning forest* F of non-connected graph with *r* components is the set of the *r* spanning trees, one for

Figure 2.11: A forest (source: Bollobas, 2001)

each component. In general, when we deal with unconnected graphs we will speak of forests and spanning forests rather than trees and spanning trees. The number of branches in a spanning forest of a graph *G* with *r* components, *N* nodes and *K* edges is equal to $b_F = N - r$ and the number of chords $h_F = K - N + r$. Therefore in the case of connected graphs (with $r = 1$), the number of branches in a spanning tree number is equal to $b_T = N - 1$ and the number of chords of the related cospanning tree is $h_T = K - N + 1$.

2.4 Basic graph-associated matrices

We saw that a graph can be represented by specifiyng the content of both sets *V* and *E* and sometimes it is also convenient to draw a diagram of the graph (see fig.2.1).

However, for more advanced purposes it is necessary to consider a matricial representation of a graph. Moreover, the analysis of the graph associated matrices is a very important part of the present work, thus we will return several times on the issue of matrix definition on a graph.

A graph $G = (V, E)$ can be completely described by giving its *adjacency matrix* (or connectivity matrix), a $N \times N$ square matrix whose entries a_{ij} with (i,j =1,...,N) are equal to 1 when the edge e_{ij} exists (when the ordered or unordered couple e_{ij} belongs to the edgeset *E*), and zero otherwise. In ordinary graph, i.e. without loops and/or multiple edges, the diagonal of the adjacency matrix contains only zeros. Moreover, **A** is a symmetric matrix for undirected graphs (asymmetric otherwise).

See the table 2.1 for an example of adjacency matrix based on the graph $G(V, E)$

(of order $N = 9$ and size $K = 18$) in fig.2.1: in this case, the adjacency matrix of *G* is a matrix $\mathbf{A}_9$ of order $n = 9$ (equal to the cardinality of the nodeset *V*).

Table 2.1: Adjacency matrix of the undirected graph in fig.2.1.

Nodes	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
v_1	0	1	0	0	0	1	1	1	0
v_2	1	0	1	0	0	0	1	0	1
v_3	0	1	0	1	0	0	0	1	1
v_4	0	0	1	0	1	0	1	1	0
v_5	0	0	0	1	0	1	1	0	1
v_6	1	0	0	0	1	0	0	1	1
v_7	1	1	0	1	1	0	0	0	0
v_8	1	0	1	1	0	1	0	0	0
v_9	0	1	1	0	1	1	0	0	0

Another undirected graph associated matrix is the *incidence matrix* **B**, a $N \times K$ matrix whose entries $B(G) = [b_{ik}]$ represent, in the general case, the number of times (0,1,2,...,m) that v_i and $e_{.j}$ are incident (for the definition of the incidence matrix in a digraph cfr. section 3.1.4). An example of observed incidence matrix, reported in table 2.2 is referred to the graph $G(V, E)$ in fig.2.1: it is a binary[10] rectangular matrix of order $N \times K = 9 \times 18$.

The adjacency matrix of a graph is generally considerably smaller than its incidence matrix, and it is in this form that graphs are commonly stored in computers [52]. For computational issues, we can often consider adjacency matrices as *sparse matrices*: it will be sometimes more convenient to implement algorithms and methods for sparse matrices in order to compute the graph invariant and the adjacency matrix associated indices.

[10] Though the general definition admits the existence of general rectangular matrix representing the incidence matrix **B** of a graph, in our case **B** is binary because the graph *G* of fig.2.1 does not contain multiple adges and/or loops

Table 2.2: *Incidence matrix of the undirected graph in fig.*2.1 (the matrix has been reduced respect to its original size).

Nodes×Edges	e_{12}	e_{16}	e_{17}	e_{18}	e_{23}	e_{27}	e_{29}	e_{34}	e_{38}	$\cdots$	e_{69}
v_1	1	1	1	1	0	0	0	0	0	$\cdots$	0
v_2	1	0	0	0	1	1	1	0	0	$\cdots$	0
v_3	0	0	0	0	1	0	0	1	1	$\cdots$	0
v_4	0	0	0	0	0	0	0	1	0	$\cdots$	0
v_5	0	0	0	0	0	0	0	0	0	$\cdots$	0
v_6	0	1	0	0	0	0	0	0	0	$\cdots$	0
v_7	0	0	1	0	0	1	0	0	0	$\cdots$	0
v_8	0	0	0	1	0	0	0	0	1	$\cdots$	0
v_9	0	0	0	0	0	0	1	0	0	$\cdots$	0

CHAPTER 3

Spectral graph theory

In this chapter we will face the major connections between vector spaces and graph. The topics in this chapter are fundamental not only for the methodological avdvances that we will present in chapter 7 but also beccause throughout it, we will described the basic elements by which it could be possible to mathematically relate the linear algebra and the multivariate statistical methods, from one hand, to the networks, from the other hand. Recalling the quotation in the beginning of Lebart, in this chapter we will illustrate the most important advanced development based on the relationship between linear algebra and graphs, namely the spectral graph theory, and its useful applications to the network science. In particular we will show how it is possible to define a very useful node-distance based on the quantity defined in the framework of the spectral graph theory [22].

3.1 Linear algebra and graphs

Let consider an undirected graph $G(V, E)$ having $N = n$ nodes and $K = m$ edges, $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$: we can define the *node-space* (or vertex-space) $\nu(G)$ of G as the vector space, over the 2-elements field $\mathbb{F}_2 = \{0, 1\}$[1]

[1] In continuous vector space the support fields are usually the set of real numbers or the set $\mathbb{C}$ of complex numbers, here we assume as support field the field $\mathbb{F}_2$ of the positive integers modulo 2.

of all functions $V \to \mathbb{F}_2$, under the operation $\oplus$ called *symmetric difference*[2]. Every element of $\nu(G)$ corresponds naturally to a subset of *V*, the subset of those vertices to which it is assigned a 1: indeed, every subset of *V* is uniquely represented in $\nu(G)$ by its indicator function. In particular, every subset V' of *V* can be represented by using a *binary n-vector* in which the *j*-th entry is equal to 1 if and only if $v_j \in V'$. Thus, $\nu(G)$ is the power set[3] of *V*, and the zero in $\nu(G)$ is the null graph (the graph with empty nodeset $\emptyset$, cfr.). Since, $\{\{v_1\}, \{v_2\}, \ldots, \{v_n\}\}$ is the canonical basis of $\nu(G)$, we have that $dim[\nu(G)] = n$.

In the same way, the space of all the functions defined on *E* having values in $\mathbb{F}_2$, $E \to \mathbb{F}_2$, i.e. the power set of *E*[4] in which every subset is represented by its indicator function, represents the *edge space* $\epsilon(G)$ of *G* where $\{\{e_1\}, \{e_2\}, \ldots, \{e_m\}\}$ is the canonical basis of $\varepsilon(G)$ (equipped with the operation of the symmetric difference) and its zero is the null graph (in particular it wold be enough a graph with empty edge set, cfr. section 3.1). Hence its dimension is $dim[\varepsilon(G)] = m$. Every subset E' of *E* can be represented by a *binary m-vector* in which the *i*-th component is 1 if and only if the *i*-th edge is in E'.

For example, let consider the graph in fig. 3.1, of order $N = 5$ and size $K = 8$: the binary *m*-vector, with $K = m = 8$, $(1, 0, 0, 1, 0, 0, 0, 1)$ represents[5] the edge subset $\{e_1, e_4, e_8\}$; thus, that vector will be an element of the edge vector space $\varepsilon(G)$. Exactly the same holds for the node vector space $\nu(G)$.

Remembering the discussion in the section 2.1.1, every subset E' of *E* is an edge-induced subgraph; as well as every subset V' of *V* is a node-induced subgraph. Hence, the edge vector space $\varepsilon(G)$ is the vector space of all the binary *m*-vector; equivalently, is also the vector space of all the edge-induced subgraphs. The same holds for $\nu(G)$.

[2]The symmetric difference (or ring sum) between two edge-sets E_1 and E_2, denoted as $E_1 \oplus E_2$, is set of all the edges that belong to E_1 or E_2 but not to $E_1 \cap E_2$. In particular, given two *r*-vectors $\mathbf{x} = (x_1, x_2, \ldots, x_i, \ldots, x_r)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_i, \ldots, y_r)$, their symmetric difference $\mathbf{x} \oplus \mathbf{y}$, is the vector $\mathbf{z} = (z_1, z_2, \ldots, z_i, \ldots, z_r)$ whose *i*-th element is $z_i = x_i \otimes y_i$, where $\otimes$ corresponds to the *Xor* operation in Boolean logic: i.e. $1 \otimes 0 = 0 \otimes 1 = 0$, $0 \otimes 0 = 0$, $1 \otimes 1 = 0$.

[3]The set of all possible 2^n subsets of *V*, including the empty set and the whole set *V*.

[4]The set of all possible 2^m subsets of *E*, including the empty set and the whole set *E*.

[5]Representing a subset, in this case, means that the bynary *m*-vector is the indicator function of the subset, hence the vector is an element of the image of the function $E \to F_2$

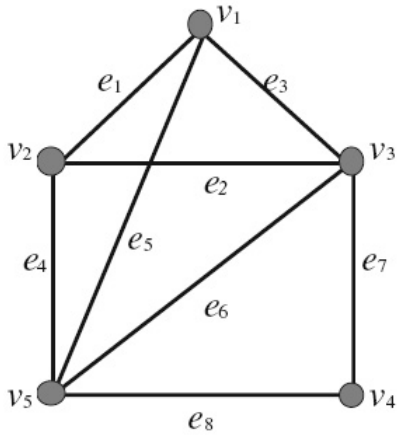


Figure 3.1: The graph G on which are based the examples in the present section.

Since the edges of a graph carry its essential structure, we shall mostly be interested in the edge space, however all the following results can be applied for the node space as well. Given two edge sets E' and E'' belonging to $\varepsilon(G)$ and their coefficients $\lambda_1^{'}, \lambda_2^{'}, \ldots, \lambda_m^{'}$ and $\lambda_1^{''}, \lambda_2^{''}, \ldots, \lambda_m^{''}$ with respect to the canonical basis, we can define their *scalar product*:

$$\langle E', E'' \rangle = \lambda_1^{'}\lambda_1^{''} + \cdots + \lambda_m^{'}\lambda_m^{'} \tag{3.1}$$

Note that $\langle E', E'' \rangle = 0$ may hold even when $E' \neq \emptyset$ or $E'' \neq \emptyset$: given two nonempty edge subsets, the scalar product between the two corresponding edge vectors is equal to zero indeed if and only if E' and E'' have an even number of edges in common. Given a subspace $\mathcal{F}$ of $\varepsilon(G)$, we can write:

$$\mathcal{F}^{\perp} = \{D \in \varepsilon(G) : D \cdot F = 0 \forall F \in \mathcal{F}\} \tag{3.2}$$

that is again a subspace of $\varepsilon(G)$ such that $dim(\mathcal{F}^{\perp}) + dim(\mathcal{F}) = m$. This holds because $\mathcal{F}^{\perp}$ is the kernel of the bilinear map $\langle \cdot, \cdot \rangle$.

We illustrated the important concept of *orthogonal subspaces*: from linear algebra

we know that two subspaces W' and W'' are said to be *orthogonal* if their scalar product is equal to zero for each pairs of vector belonging to W' and W''. It will turn back later on this important concept in section 3.1.3.

Obviously, given that the nodeset and the edgeset can be considered as linear spaces, the graph-related matrices (cfr. section 2.4) can be regarded as matrices associated to linear maps between these spaces. In particular, we can affirm the following facts. Let $\mathbf{B}^t$ the transpose of the incidence matrix $\mathbf{B}$: then $\mathbf{B}$ and $\mathbf{B}^t$ define the two linear maps $\mathbf{B} : \varepsilon(G) \rightarrow \nu(G)$ and $\mathbf{B}^t : \nu(G) \rightarrow \varepsilon(G)$ with respect to the standard bases of the two vector spaces [28]. Denoting with $\mathbf{D}$ the $n \times n$ diagonal matrix in which the (i,i)-th diagonal entry is the degree d_{ii} of the i-th node (and there are zero elsewhere), it is possible to estabilish the following connection between the matrix $\mathbf{B}$ and the adjacency matrix $\mathbf{A}$ of a graph G:

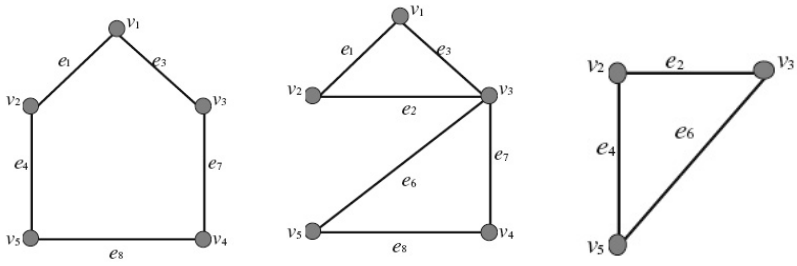
$$\mathbf{B}\mathbf{B}^t = \mathbf{A} + \mathbf{D} \tag{3.3}$$

3.1.1 Cycle subspace of an undirected graph

A graph (or a subgraph) is said to be *even* if and only if every node has an even degree. In particular, a cycle C is an even graph (or subgraph) and the null graph is considered as a cycle. We denote the set composed of all the cycles (including the null graph $\emptyset$) and all the unions of edge-disjoint cycles in a graph G[6] by $\hat{C}(G)$ with elements c. Considering these facts: (a) the edge-set E of an even graph can be partitioned into edge-subset, such that each subset in the partition forms a cycle; (b) the symmetric difference of two even graphs is always an even graph ($\hat{C}(G)$ is closed under the symmetric difference). The statements (a) and (b) allow us to affirm that $\hat{C}(G)$ is a subspace of the vector space $\varepsilon(G)$ (together with the null graph $\emptyset$)called *cycle subspace* of G. The dimension of the cycle subspace is the so-called *cyclomatic number* of a graph. Consequently an edge set E' lies in $\hat{C}(G)$ if and only if every node incident with the edges in E' has even degree.

For example in fig. 3.2 we illustrate the concept of element c of $\hat{C}(G)$, with respect

[6] A subgraph of a graph G is a cycle if and only if it is even.



(a) $c_1 \in \hat{C}(G)$ (b) $c_2 \in \hat{C}(G)$ (c) $(c_1 \oplus c_2) \in \hat{C}(G)$

Figure 3.2: (a) An element c_1 of $\hat{C}(G)$; (b) another element c_2 of $\hat{C}(G)$; (c) the symmetric difference between the two elements which is still a cycle, i.e. an element of $\hat{C}(G)$.

the graph *G* in fig. 3.1, showing how the subspace $\hat{C}(G)$ is equipped of the symetric difference as its inner operation.

By the definition of spanning tree *T* (cfr. section 2.3.2) of a graph *G*, it is easy to show that adding a chord h to *T* we obtain an unique cycle called the *fundamental cycle* C_h of *G* with respect the chord *h*. If a cycle (or generally an element *c* of $\hat{C}(G)$) contains the chords $\{h_a, h_b, \ldots, h_k\}$, then the cycle *c* can be expressed in terms of the symmetric difference between the fundamental cycles $C_a, C_b, \ldots, C_k$.

In general, given a connected undirected graph *G* and a related spanning tree *T*, there exist $m - n + 1$ different fundamental cycles (one for each chord of the cospanning tree $\bar{T}$), where *m* is the number of edges in the graph and *n* is the order of the graph. These $m - n + 1$ fundamental cycles are linearly independents in the cycles subspace $\hat{C}(G)$ and they represent a basis of $\hat{C}(G)$; consequently $dim[\hat{C}(G)] = m - n + 1$ which is also the *nullity*[7] $null(G)$ of the graph *G*. Remembering that the incidence matrix $\mathbf{B}$ of *G* can be considered as a linear map from the edge vector space to the node vector space, the nullity of *G* is the dimension of the null space of the oriented incidence matrix $\mathbf{B}$ of *G*[8]. In general, if a graph *G* consists of *p* components there are

[7] In general, the nullity of a linear map $f : V \to W$ of vector spaces is the dimension of its kernel or null space.

[8] It is often useful to give an "orientation" to the incidence matrix especially for digraphs. It is enough to specify an arbitrary orientation for the edges of *G*, in such a way that the *j*-th edge directed in the

$m - n + p$ fundamental cycles, which implies that the dimension of the kernel of G is $dim\,[null(G)] = m - n + p$.
As example, let consider again the graph G in fig. 3.1: the set of fundamental cycles with respect to the spanning tree $T = \{e_1, e_2, e_4, e_7\}$ is reported in table 3.1.

Table 3.1: Chords and corresponding fundamental cycles of the undirected graph in fig.3.1.

Chords	*Fundamental cycles*
e_3	$C_3 = \{e_3, e_1, e_2\}$
e_5	$C_5 = \{e_5, e_1, e_4\}$
e_6	$C_6 = \{e_6, e_2, e_4\}$
e_8	$C_8 = \{e_8, e_2, e_4, e_7\}$

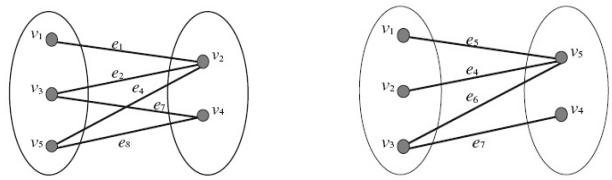
Moreover, it is easy to verify for instance that the cycle $(e_1, e_4, e_5, e_6, e_7, e_8)$ corresponding to the binary vector $(1, 0, 0, 1, 1, 1, 1, 1)$, and containing the chords e_5 , e_6 and e_8, is the symmetric difference of the fundamental cycles $\{C_5, C_6, C_8\}$: $C_5 \oplus C_6 \oplus C_8$.

3.1.2 Cutset subspace of an undirected graph

Starting from the definitions given in the section 2.1.6, here we will illustrate another important vector space, the *cutset subspace*. The set of all cutsets and of the union of edge-disjoint cutsets of a graph G is called *cutset subspace*, denoted as $\lambda(G)$, and it contains the null graph $\emptyset$. Since every cuts are union of edge-disjoint cutset, $\lambda(G)$ is also the collection of all the cuts of G. Furthermore, $\lambda(G)$ is a subset of the vector space $\varepsilon(G)$. Note that, as for cycles, the symmetric difference of any two cuts of a graph is still a cut of G, which it means that $\lambda(G)$ is closed under $\oplus$.

Let consider the graph $G(V, E)$ in fig.3.1: the two disjoint subsets of V, i.e. $V_1 = \{v_1, v_3, v_5\}$ and $V_2 = \{v_2, v_4\}$ and the other disjoint subsets $V_3 = \{v_1, v_2, v_3\}$ and $V_4 = \{v_4, v_5\}$. Let consider the cuts in fig. 3.3: $S_1 = \langle V_1, V_2\rangle$ and $S_2 = \langle V_3, V_4\rangle$, therefore $S_1 = \{e_1, e_2, e_4, e_7, e_8\}$ and $S_2 = \{e_4, e_5, e_6, e_7\}$ and $S_1 \oplus S_2 = \{e_1, e_2, e_5, e_6, e_8\}$ which is again a cut of G.

opposite sense to respect the specified orientation has a negative sign in the $\mathbf{B}$, i.e. there is a -1 in the j-column of $\mathbf{B}$. The orientation of the incidence matrix shall be defined in 3.1.4



(a) $c_1 \in \hat{C}(G)$ (b) $c_2 \in \hat{C}(G)$

Figure 3.3: (a) A cut of the graph in fig. 3.1; (b) a cutset of the graph in fig. 3.1.

Let *T* be a spanning tree of a graph *G* and let *b* be a branch of *T*. Indicating with V' and V'' the nodesets of the two components of *G* obtained by considering the graph $T - b$, it is easy to verify the cut $\langle V', V'' \rangle$ is a cutset of *G*. In particular this cutset is called the *fundamental cutset* of *G* with respect to the branch *b* of *T*. In particular, given a connected graph *G* and a spanning tree *T*, there exists $n - 1$ fundamental cutsets of *G*, one for each branch of *T* (therefore, any branch of *T* generates just one fundamental cutset of *G*). These $n - 1$ cutsets are linearly independent in the cutset subspace $\lambda(G)$ and they represent a basis of $\lambda(G)$: hence the cutset subspace with respect a spanning tree *T* of a connected graph *G* has dimension $n - 1$. This quantity represent the rank of a connected graph *G*, $\rho(G) = n - 1$, i.e. the rank of the incidence matrix **B** of *G*.

In general, if the graph *G* consists of *p* components, the number of fundamental cutsets is equal to $n - p$ and we can define the rank $\rho(G)$ of a disconnected graph *G* as $\rho(G) = n - p$, which is again the rank of the related incidence matrix **B**.
An *incidence set* of a node *v* of a graph *G*, is the cut containing the edges of *G* incident with *v*. The incidence sets of any $n - 1$ nodes of a connected graph of order $N = n$ is a basis of the cutset subspace $\lambda(G)$.

3.1.3 Relationship between cycle subspace and cutset subspace of an udirected graph

In this section we will illustrate the duality between cuts and cycles in a graph. The scalar product between a cut vector (i.e. a binary *m*-vector representing a cut) and a cycle vector (i.e. a binary *m*-vector representing a cycle), over the field F_2 is zero under

the symmetric difference operation. This fact means that a cut vector and a cycle vector always, when the graph *G* is connected, have an even number of edges in common. As a consequence, the cycle space $\hat{C}(G)$ and the cut space $\lambda(G)$ of any graph satisfy:

$\lambda(G) = \hat{C}(G)^{\perp}$ and $\hat{C}(G) = \lambda(G)^{\perp}$

Indeed, the cycle subspace and the cutset subspace of a graph *G* are orthogonals to each other. Moreover, we can affirm that: a *m*-binary vector is a cycle vector if and only if it is orthogonal to every cut vector (equivalently, a subgraph of *G* belongs to the cycle subspace if and only if it has an even number of edges in common with every subgraph in the cutset subspace of *G*); on the other hand, every *m*-binary vector is a cut vector if and only if it is orthogonal to every cycle vector $c \in \hat{C}(G)$.

An important definition from general linear algebra is the one of *orthogonal complement*: in general, considering $\oplus$ as inner operation on *W*, two subspaces W' and W'' of a vector space *W* are said to be orthogonal complement if every vector in *W* can be expressed as the symmetric difference of a vector of W' and a vector of W''. The cycle and the cutset subspaces of a graph are orthogonal complements if and only if the graph has an odd number of spanning forests (cfr. section 2.3.2). In this case every subgraph (and the graph itself) can be expressed as a symmetric difference of an element of $\hat{C}(G)$ and an element of $\lambda(G)$. This definition allows us to operate a cycle-cut based decomposition of the whole graph (or of any of its subgraph).

3.1.4 Cycle and cutset matrices for generalized graphs

If G is a digraph, the discussion above is valid with the only difference of the orientation. The cycles, the cuts and the spanning trees of a digraph *G* exactly correspond to the same quantities defined in the underlying undirected graph of *G*. However, the ways of traversing a cycle or a cut, called respectively *cycle orientation* and *cut orientation*, may have two directions: clockwise and counter-clockwise. Once we choose the orientation, a directed edge (v_i, v_j) in the cuts or in the cycles are said to be in agreement with the cut or cycle orientation if and only if the traversal of e_{ij} specified by that orientation is from its starting node v_i to its terminal node v_j [89]. In particular, the cycle subspace $\hat{C}(G)$ and the cut subspace $\lambda(G)$ of a digraph *G* are subspaces of the vector space $\varepsilon(G)$,

defined over the real field.

We will define now the principal matrices associated to the cuts and to the cycles in a directed graph *G* (they are consistent also with undirected graphs by simply disregarding the orientation). Let *G* be a digraph of size $K = m$ and of order $N = n$, having as edge set $E = \{e_1, \ldots, e_m\}$ and as nodeset $V = \{v_1, \ldots, v_n\}$. Let *C* be a cycle in *G* with a specified orientation. The cycle vector representing *C* is the *m*-vector $(x_1, x_2, \ldots, x_m)$, in which[9]:

$$x_i = \begin{cases} 1, & \text{if } e_i \in C \text{ agrees with the orientation of } C \\ -1, & \text{if } e_i \in C \text{ does not agree with the orientation of } C \\ 0, & \text{if } e_i \notin C \end{cases}$$

Let *G* be a digraph with edge set $E = \{e_1, \ldots, e_m\}$ and node set $V = \{v_1, \ldots, v_n\}$. Let *S* be a cut in *G* with a specified orientation. The cut vector representing *S* is the *m*-vector $(y_1, y_2, \ldots, y_m)$, where:

$$y_i = \begin{cases} 1, & \text{if } e_i \in S \text{ agrees with the orientation of } S \\ -1, & \text{if } e_i \in S \text{ does not agree with the orientation of } S \\ 0, & \text{if } e_i \notin S \end{cases}$$

In general if C_i for $i = 1, \ldots, r$ are the cycles in *G* and S_j for $j = 1, \ldots, t$ are the cuts in *G*, we can define: the *cycle matrix* $\mathbf{C}$ of *G* as the $r \times m$ matrix where the *i*-th row is the cycle vector representing the cycle C_i; and the *cut matrix* $\mathbf{Q}$ of *G* as the matrix whose *j*-th row is the cut vector representing the cut S_j. The columns of the cut matrix are linearly independent if and only if they correpond to the branches of a tree; similarly, the columns of the cycle matrix are linearly independent if and only if they correpond to the chords of a cospanning tree.

Let *T* be a spanning tree of a connected digraph, the *fundamental cycle matrix* of the graph with respect to *T*, denoted by $\mathbf{C}_f$, is the submatrix of $\mathbf{C}$ whose $n - m + 1$ rows are the fundamental cycles vectors. Again, the *fundamental cutset matrix* with respect

[9] According with the previous definitions, when we consider these matrices in the undirected graphs case, the entries with -1 are not defined, because of the absence of the orientation.

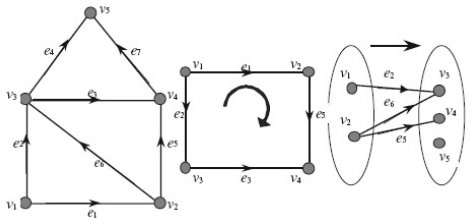


Figure 3.4: The digraph G with an oriented cycle and an oriented cut.

to *T*, denoted by $\mathbf{Q}_f$, is the submatrix of $\mathbf{Q}$ whose $n-1$ rows are the fundamental cutset vectors. The *incidence matrix* of a digraph *G* is the *n*-rowed submatrix of the cut matrix $\mathbf{B}_d$ whose rows are the incidence vectors[10] of *G*. The submatrix of the incidence matrix of a digraph *G* containing any $n-1$ of the incidence vectors is called *reduced incidence matrix* and it is denoted as $\mathbf{B}$.

Let consider the digraph in fig. 3.4: the corresponding cycle and cut vectors, with respect the chosen orientations, are respectively the vectors: $(1-1-10100)$ and (0100110). The same adjacency-incidence matrix relationship obtained for undirected graphs (cfr. the equation 3.3) is valid for a digraph $G(V,E)$ as well:

$$\mathbf{B}_d\mathbf{B}_d^t = \mathbf{D} - \mathbf{A} \tag{3.4}$$

However, this relationship in the directed case generates a fundamental graph-related matrix called *laplacian matrix* and denoted as $\mathbf{L}$ (see the section **??** for the analysis of a number of properties of the Laplacian of a graph). The elements l_{ij} of $\mathbf{L}$ are:

$$l_{ij} = \begin{cases} d_{ij}, & \text{if} i = j \\ -1, & \text{if } (v_i, v_j \in V \\ 0, & \text{otherwise} \end{cases}$$

Where d_i is the degree of the *i*-th node. In practice, the diagonal element (*i*,*i*) of $\mathbf{L}$ corresponds to the degree of the node v_i and the off-diagonal element (*i*,*j*) of $\mathbf{L}$

[10]The incidence vectors are the vectors representing the incidence set of the nodes of a digraph *G*.

corresponds to the negative of the number of the egdes connecting v_i and v_j, regardless the orientations of these edges.
One of the most important result in algebraic graph theory is related to the laplacian matrix: the *matrix-tree theorem* (also known as Kirchoff's theorem). The matrix tree theorem affirm that for a connected graph *G*, every *cofactor*[11] of the matrix $\mathbf{B}_d\mathbf{B}_d^t$ equals the number of spanning tree of *G*.

3.2 Elements of Spectral Graph Theory

This section deals with the major results of an important branch of graph theory: the *spectral graph theory* [22]. We will show how this particular field of study represents the main *trait d'union* between graph theory and linear algebra. In particular we will fully use the conlusions showed in section 3.1.

Spectral graph theory is the study of the graph associated eigenspaces. The study of the graph spectra is related to the matrix we use to refer to the graph. One can consider the classical adjacency matrix but also different associated matrices. We saw in sect. 2.4 and in sect. 3.1 several of such matrices. Some of them they were very straigthforward (adjacency and incidence matrices have an obvious meaning), other they were a litte bit more complicated (for example, cutset and cycle matrices). However we did not directly introduced the most important matrix for the study of the graph eigenspaces. At the end of the previous sect. 3.1 we introduce in Eq. the relationship between the adjacency and the incidence matrix. That particular equation define the socalled Laplacian of a graph.

We begin this section by summarizing previously known results about the Laplacian of a graph, introducing the normalized Laplacian and Fan Chung's directed Laplacian. We continue the section with a proof of the directed Cheeger bounds for the directed Laplacian by reducing to the undirected case. Throughout this section we will discuss some nice result in application of spectral graph theory and its possible implementation in SNA.

[11] A cofactor of a matrix **A** is the determinant of some smaller square matrix obtained from **A** simply removing one or more of its rows or columns.

3.2.1 The Laplacian matrix of a Graph

The *laplacian matrix* is formally the discrete analogue of the *Laplace-Beltrami operator*. Indeed historically one of the main motivations for the intersest toward the graph eigenvalues was the study of vibrations of memebranes. These study involve partial derivatives equantions. In particular, by discretization of the membranes into a grid of particles lead to the study of the matrix related to the grid. It is possible to demonstrate that the eigenfunction of a particular matrix called laplacian is the solution of the differenial equation.

Laplacian matrix $\mathbf{L}$ has several other useful properties (cfr. [15]). In the spectral graph theory literature, there are several definitions of the laplacian [22]. Unless stated otherwise, we will use the unnormalized version of $\mathbf{L}$ (also known as combinatorial laplacian.):

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{3.5}$$

Therefore $\mathbf{L}$ will have the same diagonal elements of $\mathbf{D}$ and the opposite of the elements of $\mathbf{A}$.

We do not go into detailed demonstration of laplacian properties, however we will mention the one that according to our opinion could most useful in the study of real networks. In particular we will apply some of these laplacian features to the study of different problem in SNA (in particular the ones related to the distances among actors and networks itself, cfr. chapter 7)

The first and most well known property of $\mathbf{L}$ consists in the socalled Matrix-Tree theorem, due to Kirchoff, related to the number of spanning trees in a graph (see sect. 2.3.2). This theorem states that the cofactors of $\mathbf{L}$ give the number of spanning trees in the graph. Denoting with $\mathbf{L}_{[i,j]}$ the submatrix of $\mathbf{L}$ with the *i*-th row and the *j*-th column removed it is possible to demonstrate that:

$$T(G) = (-1)^{i+j} det(\mathbf{L}_{[i,j]}) \tag{3.6}$$

where $T(G)$ is the number of spanning trees of G.

The prove of Eq. 3.7 is easily obtained by showing that all the cofactors are equal [15]. Considering now the main features of $\mathbf{L}$, namely its spectra, it is also possible to show that, by looking at linear coefficients of its characteristic polynomial, the product of all its eigenvalues λ_i with $i = 2, \ldots, n$ is n times the number of spanning trees:

$$\prod_{i=2}^{n} \lambda_i = nT(G) \tag{3.7}$$

It is worth to note that the smallest eigenvalue is alway zero $\lambda_1 = 0$ and that the second small eigenvalue λ_2 is zero *if and only if the graph is disconnected*, i.e. there are more than one component.

The importance of the eigenvalue λ_2 of $\mathbf{L}$ consists therefore in the fact that is closely related to the connectedness of the graph. Roughly speaking, large values of λ_2 are associated with graphs that are hard to disconnect. This property directly suggest the fact that the spctra of $\mathbf{L}$ can be used to find cohesive subgraphs by finding large λ_2. Some fixed values of λ_2 for particular kind of graphs are reported in the table 3.2. From table 3.2 it is clearly possible to note that for more connected graphsthe value of λ_2 is larger.

Another important result related to the eigenspace of $\mathbf{L}$ concerns the multiplicity of the null eigenvalue λ_1. In particular the number of connected components of G is equal to the multiplicity of the eigenvalue 0.

Table 3.2: Values of λ_2 of $\mathbf{L}$ for different kind of graphs on n nodes and m edges.

Chords	*Fundamental cycles*
Path	$P_n \lambda_2 = 2(1 - cos\pi/n)$
Cycle	$C_n \lambda_2 = 2(1 - cos(2\pi/n))$
Cube	$Q_m \lambda_2 = 2$
Complete	$K_n \lambda_2 = n$
Complete Bipartite	$K_{n_1,n_2} \lambda_2 = min\{n_1, n_2\}$
Star	$S_n = K_{1,n-1} \lambda_2 = 1$

There exist serveral other connections between laplacian eignevalues and graph parameters described in chapter 2. One of the most useful is the relation with diameter

of a graph. It is important because this porperty concerns with node distances. In particular, the following inequality holds:

$$\lambda_2 \geq \frac{1}{nDiam(G)} \tag{3.8}$$

3.2.2 The Laplacian pseudo-inverse

The most important results for our purposes (cfr. sect. 7.1) concerns not directly the laplacian but its pseudo-inverse. For connected graphs $\mathbf{L}$ has not full rank, therefore to obtain its inverse we have to compute its generalized inverse or Moore-Penrose pseudo-inverse [11] indicates with $\mathbf{L}^{+}_{MP}$:

$$\mathbf{L}^{+}_{MP} = \left(\mathbf{L} - \frac{\mathbf{e}\mathbf{e}^t}{n}\right)^{-1} - \left(\frac{\mathbf{e}\mathbf{e}^t}{n}\right) \tag{3.9}$$

where: n is the number of nodes and $\mathbf{e}$ is the unit column vector. This matrix in Eq. 3.9 is not the only possible inverse of the laplacian. Another way to invert $\mathbf{L}$ is by using the following formula:

$$\mathbf{L}^{+} = (\mathbf{I} + \mathbf{L})^{-1} \tag{3.10}$$

The advantage of the Eq. 3.10 to respect the Eq. 3.9 in inverting $\mathbf{L}$ is that this version of the pseudo-inverse works also in case of zero rows or column (i.e. in case in the garph there are isolated nodes).

The matrix $\mathbf{L}^{+}$ contains some interesting properties of the graph. Indeed the eigenvalue of $\mathbf{L}^{+}$ are related with the ones of the laplacian:

$$1 =$$

CHAPTER 4

Complex Networks and Network Theory

This chapter deals with the most important concepts used in complex network analysis and in social network analysis. We shall try to uniform the terms in such a way concepts basically used in complex networks theory could be adopted in describing social networks as well.

In the following we will consider networks where the vertices and edges are physical objects. Like the network of roads or railroads or the brain synapses of the nervous system in the human body. For these networks the number of edges to a vertex is limited by the physical space.

The present chapter contains a brief description of the most recent advances in the characterization and modeling of the *structural properties* of a network. In particular, basic notations and the most important definitions in network topology will be introduced. Most of this terminology coincides with the notations normally used in graph theory (cfr. chapter 2).

In this paragraph we shall describe the basic concepts, the statistics and the most applied indices used in the study of the complex networks. These quantities arose in the complex network theory as a consequence of the applications of the graph theory concepts.

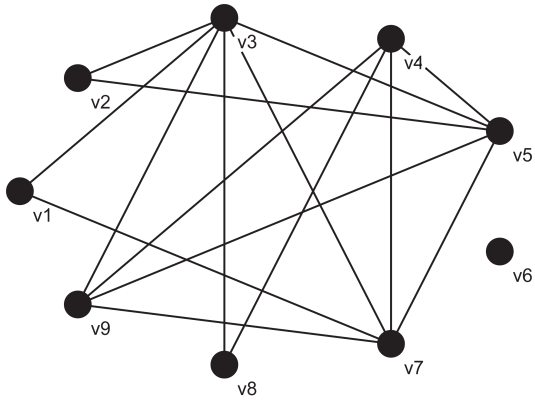


Figure 4.1: A Random graph (see section 5.1) on 9 nodes on which some vertex related quantity shall be computed.

4.1 The structural indices of the complex networks

The network studies genarally concern the network-forming forces. Such studies can be qualitative: if for example a large part of a social structure consists of isolated subnetworks, then we can conclude that the society is very segregated. We can also try to quantify the network-forming forces from the structure of the network: given a social structure we can estimate how strong the segregation in society is by measuring how many links have to be cut to split the network into strong connected communities.

In many economical and biological networks the individual vertices have different roles and functions. Another approach to network structure is therefore to label and classify the vertices with respect to each other by means of structural indices such the ones presented in the following.

4.1.1 Reachability

One of the most important concept, that has been studied since the first appearance of graph theory, is the one of *reachability* of two different nodes of a graph. This concept is strictly related to the one of distance between pairs of nodes (cfr. sections 2.1.2 and 2.1.4). Indeed, two nodes that are not connected by an edge may never be

Table 4.1: Some vertex quantities related to the random graph in 4.1: degree d_i and normalized degree (see sect. 4.1.2), closeness centrality $C_c(i)$ (see sect. 4.1.4), betweenness centrality $C_B(i)$ (see sect. 4.1.6) and the local clustering coefficient Γ_i (see sect. 4.1.7). In almost all networks, the centrality measures are, on average, quite correlated. However, the single nodes might be central in one measure and peripheral in another. It is worth to note the indices for the isolated node v_6 (the indices have been computed by using the software Pajek 1.21 - Copyright (c) 1996, V. Batagelj and A. Mrvar).

Nodes×Measures	d_i	$norm(d_i)$	$C_c(i)$	$C_B(i)$	Γ_i
v_1	0.25	1	0.518	0	1
v_2	0.25	0	0.518	0	1
v_3	0.75	0	0.778	0.232	0.333
v_4	0.50	0	0.622	0.053	0.5
v_5	0.62	0	0.691	0.08	0.6
v_6	0	1	0	0	0
v_7	0.62	0	0.691	0.08	0.6
v_8	0.25	0	0.518	0.009	0
v_9	0.50	0	0.622	0.009	0.833

reachable from one to the other. Moreover, given this unconnectedness, their distance is infinite. As we have already seen, reachability is based on several structural quantities as: *walk*, *trail*, *path* and *cycles*. We also defined geodesic, as the metric induced on a graph.

Reachability is also based on the concept of graph connectivity related to the analysis of the graph *components* (cfr. sections 2.1.3 and 2.1.5). The reachability defined on graph components is fundamental in connectivity-based network disciplines as SNA. Indeed, in SNA the components represent groups of individuals and a fundamental issue is to study how these groups are connected within each other and how they are related through the network (see Chapter 6).

4.1.2 Degree distribution

We mentioned above that reachability can be interpret in terms of connectivity. However, the concept of connectivity is a consequence of the reachability of pairs of nodes in the network structure. In particular, we define the connectivity as a measure of how the nodes are reachable from one to another. The basic quantity on which the graph connectivity is based is the *degree* of a node *i* denoted as d_i or $deg(i)$ (also known

simply as *connectivity* of a node). The node degree is the number of edges incident with that node (loops are counted twice), and is computed using the elements of the adjacency matrix **A** (cfr. section 2.4):

$$d_i = \sum_{j \in N} a_{ij} \tag{4.0}$$

If the graph is directed, the degree of the node has two components: the *out-degree* the number of outgoing edges $d_i^{out} = \sum_{j \in V} a_{ij}$, and the *in-degree*, the number of ingoing edges $d_i^{in} = \sum_{j \in V} a_{ji}$. The *total degree* is then defined as $d_i = d_i^{in} + d_i^{out}$. The following result is true [28].

LEMMA 1: *If v is a node of a graph G and* $f : G \rightarrow G^f$ *is an automorphism of G then the image w of v:* $w = v^f$*, by means of f has the same degree of v, i.e. the degrees are invariant to the automorphism.*

Proof. Let $G(v)$ the subgraph of G induced by the neighbours of $v \in V$. Then, following from the definition of automorphism we have:
$G(v)^f = G\left(v^f\right) = G(w)$ because $w = v^f$. Therefore, $G(v) \cong G(w)$ are isomorphic subgraphs of G. Consequently, from the definition of graph isomorphism, we know that $G(v)$ and $G(w)$ have the same number of nodes and v and w have the same degree. □

A list of the degrees of each node in a graph, listed in a non-decreasing order, is called the *degree sequence* of G. Thus, the degree sequence of G is a finite, non-decreasing sequence of nonnegative integer numbers whose sum is always even[1]. Conversely, given any non-decreasing sequence of nonnegative integers whose sum is even, it represents a degree sequence of a graph (but not necessary of a canonical graph, because it may contain loops and/or multiple edges). Formally we can denote a degree sequence as an ordered set $D = (d_1, d_2, \ldots, d_N)$ such that $\sum_{i=1}^{N} d_i = 2K$.

Given a degree sequence, it is very useful to think about it in terms of its frequency distribution among the N=n nodes of G. This idea is the basis of a number of conclusions on the characterisation and modelling of a graph. Indeed, the most basic topological

[1] The sum of all the degrees is always twice the number of the edges in a graph (see Eq. 4.1.2).

characterisation of a graph *G* can be obtained in terms of the related *degree distribution* p_d or $p(d)$. The *degree distribution* is defined as the probability that a random chosen node has degree d or, equivalently, as the relative frequency of nodes in the graph having degree *d*. Alternatively, the degree distribution is denoted as p_d, to indicate that the variable d assumes non-negative integer values. A plot of p_d for any given network can be depicted by making a histogram of the degrees of vertices. This histogram graphically represents the degree distribution for the network. The important notion here is that the degree distribution summarize the basical topological properties of a network. For this reason this distribution represents the most important graph characteristic in the analysis of network structure.

In the case of directed networks one needs to consider two distributions, $p\left(d^{out}\right)$ and $p\left(d^{in}\right)$. Also for bipartite graphs (cfr. sec.1) one has to consider two degree distributions, one for each subset of the nodes involved in that network.

The distribution p_d is the usual probability distribution of a discrete random variable d. Indeed, information on how the degrees are distributed among the nodes of an undirected graph[2] can be obtained either by a plot of p_d or by the corresponding cumulative node degree distribution function (CDF) P_d:

$$P_d = pr(d \geq d') = \sum_{d'=d}^{\infty} p_{d'} \tag{4.0}$$

which is the probability of observing a node degree equal or greater than *d*.

The degree is a quantity that has been investigated very intensively in recent years and this interest is due to a discovery, of Barabasi et al. in [45] and [4], relative to the probability density function of the degree. In particular it has been proved that many real-world networks have a power-law degree distribution (see section 5.5). For standard random variables one can compute the *r*-th moments and the *s*-th central moments of the distribution. The *r*-moment of P(*d*) (with $r = 1, 2, \ldots$) is defined as:

[2] We dealing with the case of undirected network. In the digraph it is sufficient to consider the two distributions of in-degree and out-degree. However the considerations will be exactly the same.

$$\bar{d^r} = \sum_d d^r P(d) \tag{4.0}$$

The first moment $\bar{d}$ is the mean degree of *G* (also denoted with z in the following) and it is also computed by using the size *K* and the order *N* of the graph:

$$\bar{d} = 2K/N \tag{4.0}$$

The second moment $\bar{d^2}$ (or $\bar{d}^2$) measures the fluctuations of the connectivity distribution, and, as we shall see in the following, the divergence of $\bar{d^2}$ in the limit of infinite graph size, radically changes the behavior of dynamical processes that take place over the graph.

We can define as well the central moments of the degree distribution:

$$\bar{d_c^r} = \sum_d \left(d - \bar{d}\right)^r P(d) \tag{4.0}$$

The degree distribution completely determines the statistical properties of *uncorrelated networks*. However, it is worth to note that usually, the real networks have a strong correlated structure because the probability that a node *i* of degree d_i is connected to another node *j* of degree d_j depends on d_i.

In these cases, it is necessary introduce a conditional probability degree distribution $P(d_j|d_i)$, being defined as the probability that an edge from a node of degree d_i points to a node of degree d_j.

Indicating d_i as *d* and d_j as *d'*, it is possible to affirm that $P(d'|d)$ satisfies the normalization $\sum_{d'} P(d'|d=1)$, and also the degree detailed balance condition $dP(d'|d)\,P(d) = d'P(d|d')\,P(d')$ [13]. For uncorrelated graphs, in which P(*d'*|*d*) does not depend on *d*, the detailed balance condition and the normalization give $P(d'|d) = \frac{d'P(d')}{\langle d \rangle}$.

Although the degree correlations are formally denoted by $P(d'|d)$, the direct computation of the conditional probability often gives very noisy results for most of the real networks because of their finite size *N*. This problem can be overcome by defining the average nearest neighbors degree of a node *i* as:

$$d_{nn,i} = \frac{1}{d_i} \sum_{j \in V_i} d_j = \frac{1}{d_i} \sum_{j=1}^{N} a_{ij} d_j \tag{4.0}$$

where the sum varies on the nodes belonging to i V_i , the set of the neighbors of i. By using the definition 4.1.2, it is possible to compute the average degree of the nearest neighbors of nodes with degree d, denoted as $d_{nn}(d)$, obtaining an expression that implicitly incorporates the dependence on d [68]. This quantity can be expressed in terms of the conditional probability as:

$$d_{nn}(d) = \sum_{d'} d' P(d'|d) \tag{4.0}$$

If there are no degree correlations, formula 4.1.2 becomes $d_{nn}(d) = \langle d^2 \rangle / \langle d \rangle$, which it means that $d_{nn}(d)$ is independent from d. Correlated graphs are called *assortative* if $d_{nn}(d)$ is an increasing function of d, whereas they are referred to as *disassortative* when $d_{nn}(d)$ is a decreasing function of d [**?**]. In particular, in assortative networks the nodes in the network that have many connections tend to be connected to other nodes with many connections, while in disassortative networks nodes with low degree are more likely connected with highly connected ones. Social networks are in general classified as assortative whereas technological networks are mostly of disassortative type.

Degree correlation are usually quantified by reporting the numerical value of the slope of $d_{nn}(d)$ as a function of d, denoted in the following as ν, or by calculating the Pearson correlation coefficient of the degrees at either ends of an edge [**?**]. In the following we shall meet some real networks showing both types of correlations.

4.1.3 Density

The most simple structural characteristic in a graph is the *density*. While degree is a concept involving the number of edges incident with the individual nodes in a network, with the density we consider the number (and the proportion) of the edges in the whole graph. This concept is defined both in deterministic and stochastic terms. In the former

definition we have to distinguish between simple and undirected graphs. In general a graph can only admit a given number of lines, depending on the number *N* of nodes (in practice, the size depends on the order). As we know (cfr. section 3.1), given $N = n$ nodes in an undirected graph without loops and multiple edges, the maximum number of possible lines in *E* (i.e. the size *K* of the graph or cardinality $|E|$ of *E*) is equal to the number of combination of the nodes considered pairwise:

$$Max(K) = \left(\binom{n}{2}\right) = \frac{n(n-1)}{2} \tag{4.0}$$

Given this theoretical maximum one can consider the proportion of the observed edge *K* in a graph *G* (with $N = n$) respect with its maximum size $Max(K)$. This ratio is the density δ of *G*:

$$\delta = \frac{K}{Max(K)} = \frac{K}{n(n-1)/2} = \frac{2K}{n(n-1)} \tag{4.0}$$

The density of a graph varies from $\delta = 0$ if $K = 0$ i.e. we have an empty graph, to $\delta = 1$ if *K* assumes its maximum value $K_{max} = n(n-1)/2$, i.e. we observe the complete graph. It is worth to note that the complete graph contains all possible edges and all node-degrees are equal to $d_i = n - 1 \; \forall i \in V$, i.e. the node distribution is distributed as an uniform distribution.
Thus δ is a relative measure of how a given graph *G* is far from the completness. There is also a direct relationship between the density and the first moment of the degree distribution of a graph *G* combining equations 4.1.2 and 4.1.3:

$$\delta = \frac{\bar{d}}{n-1} \tag{4.0}$$

Equation 4.1.3 indicates that density is the average proportion of edges incident with nodes in the graph.

4.1.4 Closeness centrality

The concept of degree is used also to indicate the centrality of a node. However, an high degree vertex could be central only in its nighbourhood, but can be peripheral in the graph as a whole, in the sense that the average distance to other vertices is large. For this reason a degree index should be followed by the closeness centrality [76]:

$$C_c(i) = \left[\sum_{i,j \in V, i \neq j} g_{i,j} \right]^{-1} \tag{4.0}$$

Eq. 4.1.4 is simply the inverse of the sum of the geodesic between the actors. Moreover $C_c(i)$ it is meaningful only for connected graphs. At its maximum this index is equal to $N-1$, a normalized version of Eq. 4.1.4 is the one proposed by Beauchamp [7]:

$$C_c^{'}(i) = \left[\frac{N-1}{\sum_{i,j \in V, i \neq j} g_{i,j}} \right] = (N-1) C_c(i) \tag{4.0}$$

The index in Eq. 4.1.4 ranges between 0 and 1. This global centrality measure is the simplest one, in the sense that the closeness centrality for an individual vertex is influenced by the wiring of the entire graph [49].

4.1.5 Eigenvectors centrality

The method called *eigenvector centrality* [17] w_v consists in assigning to the vertices $v \in V$ a positive weights w_v proportional to the sum of weights of the in-neighbourhood:

$$_v = \lambda^{-1} \sum_{u \in H_v^{in}} w_u \tag{4.0}$$

or in matricial form: $\mathbf{A}\mathbf{w} = \lambda \mathbf{w}$, where $\mathbf{A}$ is the adjacency matrix of the graph, λ is the largest eigenvalue and $\mathbf{w}$ is the corresponding normalised eigenvector (the eigenvector with unitary norm).

The eigenvector centrality measure has found an important application in the design of WWW search engines [72]. Since a hyperlink indicates that the web-page referred to has some relevance for the reader, the eigenvector centrality gives an estimate of the overall importance of the page [49]. To sort the entries to a query, search engines use eigenvector centrality.

We may obtain additional centrality information considering the eigendecomposition of different network associated matrices, rather than decompose $\mathbf{A}$. In particular, since information centrality focuses on the information contained in all paths originating when a specific actor "signals" (information flows, communication, etc.) travel along the paths between vertices, then one can assume that the information is inversely proportional to the variance of the signal strength. In chapter 3, we will show how the use of spectral decomposition of some graph associated matrices is useful to obtain different information on network structure.

4.1.6 Average shortpath length and betweeness

We already mentioned the problem of measuring distance in a connected graph by using the so-called *geodesic g* or *shortest path length* (cfr. sectionDistance in graphs).

Moreover, shortest paths have also played an important role in the characterization of the internal structure of a graph [91] [78]. It is useful to represent all the shortest path lengths of a graph *G* in a distance matrix $\mathbf{\Delta}$ in which the entries g_{ij} represent the length of the geodesic from node *i* to node *j*.

A measure of the typical separation between two nodes in the graph is given by the *average shortest path length*, also known as *characteristic path length*, defined as the mean of geodesic over all couples of nodes [95]:

$$\bar{g} = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} g_{ij} \tag{4.0}$$

A problem with this definition is that of course $\bar{g}$ it diverges if there are disconnected components in the graph (that is the reason why we specified that the graph must be a connected one).

One possibility to avoid the divergence is to limit the summation in Eq. 4.1.6 only to couples of nodes belonging to the largest connected component. An alternative approach, that can be useful, is to consider the harmonic mean [**?**] of geodesic lengths, and to define the so-called *efficiency* of G [**?**] as:

$$Eff = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} \frac{1}{g_{ij}} \tag{4.0}$$

This latter quantity can be consider an indicator of the traffic capacity of a network, and moreover it avoids the divergence of formula 4.1.6, since any couple of nodes belonging to disconnected components of the graph gives a contribution equal to zero to the summation in formula 4.1.6. The mathematical properties of the efficiency have been investigated in Criado et al [67].

The communication of two non-adjacent nodes, say *j* and *k*, depends on the nodes belonging to the paths connecting *j* and *k*. Consequently, a measure of the relevance, namely *betweenness centrality* [37], of a given node can be obtained by counting the number of geodesics going through it. The *node betweenness* is a centrality measure that capture how central a vertex is in the communication flow in networks like the Internet or social networks. Together with the degree, the betweenness is one of the standard measures of node centrality, originally introduced to quantify the importance of an individual in a social network [91]. More precisely, the betweenness $C_B(i)$ of a node *i*, sometimes referred to also as *load*, is defined as:

$$C_B(i) = \sum_{j,k \in V, j \neq k} \frac{n_{jk}(i)}{n_{jk}} \tag{4.0}$$

where n_{jk} is the number of geodesics connecting *j* and *k* and $n_{jk}(i)$ is the number of geodesics connecting *j* and *k* and passing through *i*.

The concept of betweenness can be also applied to the edges. In particular the *edge betwenness* is the number of geodesics between each pair of nodes that passing through that edge [**?**].

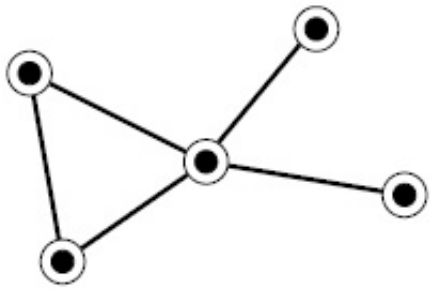


Figure 4.2: Visualization of the clustering coefficient γ, (see Eq. 4.1.7). This network has one triangle and eight connected triples, and therefore has a clustering coefficient of $3 \times \frac{1}{8} = \frac{3}{8}$. The individual vertices have local clustering coefficients Γ_i (see Eq. 4.1.7) equal to the following vector: $(1, 1, \frac{1}{6}, 0, 0)$ and a mean value Γ (see Eq. 4.1.7) equal to $\Gamma = \frac{13}{30}$ (source: M.E.J. Newman, 2003).

4.1.7 Clustering or transitivity

Clustering also known as *transitivity* (especially in social network analysis) can be simply described as the measure of the tendency of two persons to know each other given that they have a friend in common. In general, a relation *R* between three nodes *i*, *j* and *k* is transitive if iRj and iRk then jRk. In terms of general graphs, the transitivity γ indicates the presence of a certain number of triangles (remember from the section 2.1.2 that a triangle C_3 is the simplest cycle in a graph). Operatively the transitivty is measured as the proportion of connected triads that also form a triangle on the total number of triads in the graph:

$$\gamma = \frac{3 \times N_3^C}{N_3^U} \tag{4.0}$$

where: N_3^C is the number of triangles in *G* and N_3^U is the number of connected triples of vertices of *G* (triads forming or not forming a triangle); the factor 3 compensates the fact that each triangle is counted three times in the connected triples (one for each of the three nodes involved in the relation). This correction ensures that $0 \leq \gamma \leq 1$ with $\gamma = 1$ if the graph *G* is K_N (fig. 4.2 graphically illustrates the meaning of γ coefficient).

Another index of transitivity is the *clustering coefficient* c developed by Watts and

Strogatz [94]. In particular, in a graph $G(V, E)$ of order $N = n$ one can compute the clustering coefficient for an arbitrary node *i*: this quantity, denoted as Γ_i, indicates the probability of observing $a_{jk} = 1$ the (*j*,*k*) element of the adjacency matrix $\mathbf{A}$ of *G*, given that the nodes *j* and *k* are both neighbours of the node *i*. The local index Γ_i is computed considering the number of edges e_i in the subgraph G_i of the neighbours of the node *i* (cfr. section 2.1.1 for the definition of *i*-neighbours subgraph) divided by $k_i\,(k_i - 1)\,/2$ the theoretical maximum number of edges in G_i:

$$\Gamma_i = \frac{2e_i}{k_i\,(k_i - 1)} = \frac{\sum_{j,k} a_{ij}a_{jk}a_{ki}}{k_i\,(k_i - 1)} \tag{4.0}$$

This measure is the basic quantity in the definition of the Small World network model introduced in [94]: Fast alghorithm to compute Γ_i has been developed in [59]. In order to obtain the *global clustering coefficient* Γ of the whole graph *G* the average of Γ_i over all nodes $i \in V$ is computed:

$$\Gamma = \bar{\Gamma}_i = \frac{1}{n}\sum_{i \in V} \Gamma_i \tag{4.0}$$

Both quantities in Eq. 4.1.7 and Eq. 4.1.7 are build up as relative indices: $0 \leq \Gamma_i \leq 1$ and $0 \leq \Gamma \leq 1$.

As we stated before, the indices γ and Γ measure the relative presence of the number of the triangles, which is the smallest size cycle, in the graph. To generalize these quantities, several indices have been proposed in order to measure the proportion of cycles C_h of an arbitrary size *h* (cfr. [10]). Furthermore, other indices have been defined also to avoid the bias of the degree correlation [**?**].

4.1.8 Small World Property

The study of several dynamical processes over networks has pointed out the existence of shortcuts, i.e. bridges connecting different components, thus speeding up the communication within the network.

In regular hypercubic lattices on *d* dimension, the mean number of vertices one

has to pass by in order to reach an arbitrarily chosen node, grows with the lattice size as $N^{1/d}$ [12]. Conversely, in most of the real networks, despite of their large size, there is a relatively short path between any two nodes. This feature is known as the *small-world property* and is mathematically characterized by an average shortest path length $\bar{g}$, defined as in Eq. 4.1.6, that depends at most logarithmically on the network order *N*. This property was first investigated, in the social context, by Milgram in the 1960's in a series of experiments to estimate the actual number of steps in a chain of acquaintances [**?**] (see chapter 6).

In its first experiment, Milgram asked randomly selected people in Nebraska to send letters to a distant target individual in Boston, identified only by his name, occupation and rough location. The letters could only be sent to someone whom the current holder knew by first name, and who was presumably closer to the final recipient. Milgram kept track of the paths followed by the letters and of the demographic characteristics of their handlers. Although the common guess was that it might take hundreds of these steps for the letters to reach their final destination, Milgram's surprising result was that the number of links needed to reach the target person had an average value of just six. More recently, a similar experiment conducted by Dodds et al. [63] on e-mail exchanges successfully reproduced Milgram's experiment [12].

The small-world property has been observed in a variety of other real networks, including biological and technological ones [95], and is an obvious mathematical property in some network models, as random graphs. Indeed if the number of nodes within a distance *r* of a typical central node grows exponentially with *r* then the value of the average shortest path length $\bar{g}$ will increase as $logn$ [**?**]. In recent years the "small-worldliness" of a network has this more precise interpretation: networks are said to have the small-world property if the value of $\bar{g}$ scales logarithmically or slower with the nodeset cardinality *N* for fixed mean degree.

However, differently from random graphs, the small-world property in real networks is often associated with the presence of transitivity, denoted by high values of the clustering coefficient, defined as in Eq. 4.1.7. For this reason, Watts and Strogatz, in 1998, have proposed to define small-world networks as those networks having both a

small value of $\bar{g}$, like random graphs, and a high clustering coefficient Γ, like regular lattices [94].

4.1.9 Motifs

A motif *M* is a number of connected (both directed or undirected) s-node subgraphs $G(V_s, E_s)$ of a graph $G(V, E)$, with $s << n$, such that they occur in *G* in a significantly higher number with respect their occurrence in a randomized version G^{rand} of *G*.
To seek the proportion of the motifs in an observed graph $G(V, E)$ the approach is to consider the occurrence of specific s-node subgraphs *M* in *G* and in a random graph having the same structural characteristics of *G* (cfr. 5.1). We can measure the statistical significance of the value of *M* by its Z-score, using the following formula:

$$Z_M = \frac{n_M - \bar{n}_M^{rand}}{\sigma_{\bar{n}_M}^{rand}} \tag{4.0}$$

where: n_M is the number *n* of observed motifs M_s of order *s* in the graph *G*; $\bar{n}_M^{rand}$ is the mean of the number of motifs M_s occurred in the random graph and $\sigma_{\bar{n}_M}^{rand}$ is the standard deviation of of the number of M_n in the randomized version of *G*.
Though the concept of motifs was formalized by Alon et al. [**?**], the context in which the study of the motifs arose for the first time was the SNA. The reason of the importance of motifs in SNA is the historical relevance given to the *local structures* (see chapter 6 for more details) in human interactions studies: in particular, several sociologists (see for example Simmel in [79], or Breiger in [18]) studied the group interactions by through the analysis of small subgroups behaviour (involving most of the times two or three actors). This approach may be often useful but it is important to pay attention to the generalization of local network structures as global index of the general network structure. In the whole network may arise very different processes in organizing the actual net conformation. The motifs, in SNA are principally studied in form of *community structures* (also known as *cohesive subgroups*). Cohesive subgroups are subgraphs $G'(V', E')$ of $G(V, E)$ whose nodes are tightly connected, in such a way that the density within the communities is higher then the density between them. The

defintion of connectedness inside the communities is quantified in several different ways. In particular, the more strict defintion requires that the nodes of the subgraph are connected pairwise: this defintion corresponds to the concept of *clique*. Formally a clique is a maximal complete subgraph of $s \geq 3$ nodes all of which are adjacent to each other and no other nodes adjacent to all of them exist: We can relax the specification of connectedness in terms of reachability: a s-clique K_s is the maximal subgraph where the largest geodesic is bounded by s: $d(i,j) \leq s$ for $\forall i,j \in K_s$ (for more details see chapter 6). In particular 2-clique is the subgraph where all nodes need not to be adjacent but they are reachable through at most one intermediate node; 3-clique is the subgraph where all nodes are reachable trough at most 2 intermediaries, and so on [12].

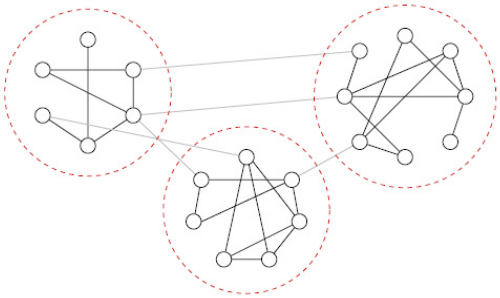


Figure 4.3: Three communities from a given connected graph. Communities are subgroups defined in such a way that the density within them are higher than the density between the subgroups (source: M.E.J. Newman and M. Girvan, 2004. © 2004 by the American Physical Society.

A different definition of community is based on density of edges within and between the community. In particular, we require that the density within the communities is greater than the density between them. An example of such definition is represented in fig.4.3.

CHAPTER 5

Models for complex networks

In this part we summarize the widely used structural network models to identify network topology. Throughout the present section, whenever we refer to an observed network G we will intend the real network under study, which it represents the concrete object of the analysis. Models is an almost omnipresent concept in the natural and social sciences. The term, in this chapter, is used in various contexts: sometimes we will refer about network models as definitions of a probability spaces, but most often we will intend generating networks schemes. In this latter meaning, a model can be trated in two differnt ways: either one looks for the minimal model generating one (or many) specific network structure, or one bases the model on measurements or observation-founded assumptions.

The traditional way of statistics (but also in the natural sciences) is to look for minimal models. Models in this sense present one single candidate for the mechanism generating the structure, and point other events of the real world situation out as irrelevant. For networks one can ask what the minimal requirements are for a certain degree distribution to emerge in a growing network.

The network models we face in this part, arise in statistical mechanics. In statistical mechanics the models are often designed in order to analyze the phase transitions in complex systems. This phenomenon is expressed as the emergence of a critical value of a model parameter that changes the system behaviour. It turns out that very different models can behave similarly in the vicinity of a critical point (in parameter space).

This phenomenon is called *universality*. Quite naturally, structural phase transitions and universality have been studied many times in network literature (see for example [**?**] and [2]). Perhaps the first example of structural phase transitions in graphs is the emergence of a *giant component*, a big cluster that contains all the vertices of *G*, as grows beyond the threshold $N/2$ (see sect. 5.1.2) in a random graph [69], [**?**].

Basically, in throughout this chapter we will discuss about groups of structural network models, precisely:

- Random Graphs and related models (Random walks, Markov chain on graph, etc.)
- Small World Networks

These kinds of models, after a period in which they have been studied as a "real" object, are actually implemeted to describe the stochastic behaviour of a given network. Specifically, by using these models we "measure", in a probabilistic fashion, how likely is to observe some given network structural property Q in *G* or, more specifically, some given network parameter $\theta(G)$ given a network topology. Here the difference between a property Q and a parameter $\theta(G)$ is a matter of convenience: we use the term parameter whenever we refer to one of the indices described in section 4.1. In general, a property can be interpret as a particular case of subgraph, and a parameter is only a particular case of graph property. Formally, given an observed structural network parameter $\theta(G) = \hat{\theta}$ (or, equivalently, a property Q) the aim is to derive the probability $P[\hat{\theta}(G^{model})]$ of observing the value $\hat{\theta}$ under a certain network generating model G^{model} (we also indicate the generating model as *probability space*), often by using with the costraint that G^{model} shares some other particular network structural property Q' with the observed real network G[1]. This procedure allows to find the probability distribution of the parameter $\theta(G)$ of interest. In the case we are interested in a particular property Q, by using these models, we can observe when, during the generating process, Q arise. The section starts with a complete description of random graph model and its characteristics, since is the basis of more recent network modelling. Subsequently we shall define some extension

[1] In the simplest case the observed network and the family of generated random graphs share just the cardinality and the labels of the nodeset.

of random graph, as generalized random graphs and included different kind of models, precisely random walks and Markov chain on graph that share some common features with random graphs. However this latter mathematical object are different to respect the other structural models, because they belong to the family of stochastic processes and they do not deal directly with network structure, but they have very nice application to graph structures (see chapter). Afterward we deal with the most recent advance of network modelling describing, in particular the Small World Network family of models and the related Scale-Free Networks.

5.1 Random Graphs

The simplest useful model defined on a network is the *random graph model* $\mathcal{G}^{rand}$, first studied by Rapoport [70], Solomonoff and Rapoport [69], Gilbert [43] and Erdos and Renyi [31]. The random graph model has been well analyzed by mathematicians, and many results, both approximate and exact, have been proved rigorously. Most of the interesting features of real-world networks that have attracted the attention of researchers in the last few years, however, concern the ways in which networks are not like random graphs, i.e. how the observerd graph properties deviate from randomness. Real networks are nonrandom in some revealing ways that suggest both possible mechanisms that could be guiding network formation, and possible ways in which we could exploit network structure to achieve certain aims. Hence, RG models are very important in exploring the real networks behaviours. For this reason we will widely illustrate the logic governing the random graphs. Following Bollobas [16], in order to study these models:

> it is often helpful to imagine a random graph as a living organism which evolves with time. It is born as a set of n isolated vertices and develops by successively acquiring edges at random. Our main aim is to determine at what stage of the evolution a particular property of the graph is likely to arise.

Formally a random graph model $\mathcal{G}$ is a stochastic process in which the starting point is a network with n vertices and no edges and at each step one new edge it is added. The

new edge is uniformly chosen from the set of missing edges. The random graph models represent the behaviour of a labelled graph of a particular type in a probability space. It is important to point out that an element of this probability space is a random graph with *n* edges and a given number of edges. We basically can consider two fundamental types of spaces. The first one is the more general space (in the undirected networks case), denoted as $\mathcal{G}(n, K)$, and it is composed of all the graphs with a given set of $N = n$ nodes and a number of edges *K* varying in the interval $0 \leq K \leq K_{max} = \binom{N}{2} = 0,5n(n-1)$ (where *K* only may assume positive integer values). Intuitively, starting with $N = n$ disconnected nodes, this probability space is generated by randomly connecting couples of selected nodes, without multiple connections, until the number of edges equals *K*. Almost always, *K* is a function of *n*: $K = K(n)$. Specifying the desired number of edges, i.e. if we set $K = m$, we can also obtain the space where both $N = n$ and $K = m$ are fixed. This space is denoted with $\mathcal{G}(n, m)$, in which every random element $G_{n,m}$ is a *m*-size graph. This model can be regarded as a "snapshot" of the evolution of random graph model $\mathcal{G}(n, K)$ made at certain *m*-th point of the process. Obviously, given that random graph model is a stochastic process, it is worth to note that at certain point of the evolution we cannot exactly determine which is the actual network state. In other words, random graphs in $\mathcal{G}(n, m)$ may represent different elements topological structures.

Obviously $\mathcal{G}(n, K)$ and $\mathcal{G}(n, m)$ differ only in their cardinality (because $\mathcal{G}(n, m)$ represents a step of $\mathcal{G}(n, K)$): for $\mathcal{G}(n, K)$ we have $2^{\binom{n}{2}}$ possible random graphs; for $\mathcal{G}(n, m)$ we have $\binom{\binom{n}{2}}{m}$ elements. It is easy to note that $2^{\binom{n}{2}} \geq \binom{\binom{n}{2}}{m}$.

To better understand the functioning of random graphs in $\mathcal{G}(n, K)$, let consider an example. Let *H* be a graph with $N = 4$ nodes. For what we affirmed above H belongs to the probability space $\mathcal{G}(N = 4, K)$ (or simply $\mathcal{G}(4)$), which is the ensemble composed of every graph with 4 nodes and a number of edges (depending on *N*) bounded between $0 \leq K \leq \binom{4}{2} = 6$. Thus the cardinality of $\mathcal{G}(N = 4)$ is $|\mathcal{G}(N = 4)| = 2^{\binom{4}{2}} = 2^6$. Therefore in this space, the probaility of observing a particular graph with 4 nodes and a given topology is equal to:

$$P(G_{n,K} = H) = 1/2^{\binom{n}{2}} \tag{5.0}$$

In our case: $P(G_{4,K} = H) = 2^{-6}$. More particularly, if *H* is an undirected graph with node set cardinality equals to $N = 4$ and $K = 2$ edge, we can specify that *H* belongs to the probability space $\mathcal{G}_{N=4,K=2}$, or simply $\mathcal{G}_{4,2}$, which is the ensemble containing every graphs with node set $N = 4$ and $K = 2$ edges (the graphs differ only by the topology determined by the labelled nodes)[2]. In particular, this space is composed of $\binom{K_{max}}{m}$ (where $K_{max} = \binom{n}{2}$) possible random graphs. In our case: $\binom{\binom{4}{2}}{2} = 15$ random graphs. Hence the probability of occurrence of the generic element *H* is equal to:

$$P(G_{n,m} = H) = 1/\binom{K_{max}}{m} \tag{5.0}$$

which is in our specific case: $P(G_{n,m} = H) = 1/\binom{\binom{n}{2}}{2} = 1/\binom{6}{2} = 1/15$.

However, it is worth to note that the most crucial point is how the edges are randomly attached to couples of selected nodes. To modelling this behaviour, we introduce a slight but theoretical important variation on our probability space (again in the undirected networks case): in particular, let consider the space $\mathcal{G}\{n, Pr(e_{ij}) = p\}$ (with $0 < p < 1$) of all the graphs with a given set of $N = n$ nodes and in which each of the $0 \leq K \leq \binom{n}{2}$ possible edges is independently[3] (or conditionally) present with some probability distribution $P_{e_{ij}}$. This kind of model, which represents a generalization of the previous probability space, is the one that we will definetly indicate as Random Graph Model (RGM from now on). Following Spencer [83] we will try to simplify it by means of a simple experiment. Let consider*n* vertices be labelled as $\{1, ..., n\}$. By flipping a coin, for each unordered pair of vertices, it is decided whether they have to be adjacents. The coin is biased to come up head with probability *p* and the edge is placed between every couple of nodes exactly when the coin comes up head. This, on an intuitive level, is the random graphs' space $\mathcal{G}\{n, P(e_{ij}) = p\}$. Formally, the

[2]The graphs in $\mathcal{G}_{n,m}$ are ISOMORPHICS in the case of unlabelled graphs; however, when we attach to each of the *n* nodes a unique label, the elements of $\mathcal{G}_{n,m}$ become different topological structures.

[3]The introduction of the independence in the probabilities of occurrence of an edge is a strong simplification: we will admit the independence only to illustrate some preliminary results.

important point is that with RGM we deal with a *finite probability space* composed of $2^{\binom{N}{2}}$ labelled graphs on $V = \{1, ..., n\}$. The probabilities are determined by assuming that $Pr[e_{ij}] = p$ and that the events $e_{ij} = \{0\ or\ 1\}$ are mutually independent. It is important to point out that $\mathcal{G}\{n, P(e_{ij}) = p\}$ is a probability space in which every singleton, and hence every set, does have a measure.

Finally, by using the above definitions we then obtain an ensemble, $\mathcal{G}\{n, P(e_{ij}) = p\}$ containing graphs with a fixed number of nodes but different number of links: graphs with *K=m* links will appear in the ensemble with a probability that follows a given distribution. We can observe that this probability distribution governing the occurrence of edges connecting couples of nodes or equivalently the number of edges attached to a given node, i.e. the degree of that node. The most used probability law is the *binomial distribution* or the *Poisson distribution* as limit distribution of $P(\cdot)$ for large *N*. Let consider an element *J* which is a graph with a given node set and $K = m$; in according with the binomial distribution we will have that the probability occurrence of *J* is equal to:

$$P(J) = P(G_{P_K \approx bin(n,p)} = J) = p^m (1-p)^{K^{max}-m} \tag{5.0}$$

Moreover, a natural refinement of $\mathcal{G}\{n, P(K) = p\}$ will be the model $\mathcal{G}\{n, p_{ij}\}$ (with $0 \leq p_{ij} \leq 1$ and $1 \leq i < j \leq n$) consisting of all graphs with vertex set $V = \{1, 2, \ldots, n\}$ in which the edges are chosen independently, and for $1 \leq i < j \leq n$ the probability of the couple $\{ij\}$[4] being an edge is exactly p_{ij}.
For practical purposes it is very useful to know that RG models $\mathcal{G}(n, K)$ and $\mathcal{G}\{n, P(K) = p\}$ are equivalent if and only if we will choose $K = Np$ which is the first moment of binomial distribution on which P_K is based on.

5.1.1 The 0-1 law.

The principal purpose of Random Graphs models is to discover if the graphs in $\mathcal{G}$ have a specific property Q: for example, the property of containing a triangle.

[4]Which is unordered, because we are still dealing with undirected graphs.

More precisely, we will say that some graph in $\mathcal{G}$ has a property Q if all the graphs in a set Q, having in common a given property Q, is a subset of a given RG family $\mathcal{G}$. Therefore, given a graph $G \in Q$ it could be stated indifferently "G has the property Q" or "G belongs to Q". Moreover, it is possible to associate Q with the *event* "the property Q holds". Hence, we shall also use the terms property or event interchangeably. Formally $Pr[G(n,p);Q]$, or simply $P_n(Q)$, is the probability that the event Q occurs in the probability space $\mathcal{G}(n,p)$. For example, if $\mathcal{G}(n,p)$ is constructed by the coin flips as described above, then $Pr[G(n,p);Q]$ represents the probability that the experiment will yield a graph with property Q. It is worth to note that, for example, in the special case $p = 0.5$, all graphs have the same probability of occurrence and so $Pr[G(n,p);Q]$ represents just the proportion $f_n(Q)$ of labelled graphs on *n* vertices having property Q. Sometimes we will use the notation $P_K(Q)$ in order to indicate that the probability of a graph of $\mathcal{G}(n,K)$ belongs to Q and the notation $P_p(Q)$ when is an element of $\mathcal{G}(n,K)$ that have Q.

According with previous definitions, we may study the occurence of a given property Q from two sources and depending on the probability space we are dealing with [83]. The first is the family of random graphs with constant probability, the space $\mathcal{G}(n,K)$; the second is the notion of evolution (or growth) of random graphs[5] as defined by Erdos and Renyi [31] and represented through the space $\mathcal{G}(n,p)$.

In that evolution it is central that the edge probability *p* be taken not as constant but as function $p(n)$ of the total number of vertices[6].

Bollobas remarks that the most important point discovered by Erdos and Renyi was that many important properties of graphs belongs to almost every graph (if we consider fixed probability spaces) or, equivalently, appears very suddenly during the random graphs growth [16]. For example, if we pick the function $K = K(n)$ then, in many cases, either almost every graph G_K in $\mathcal{G}(n,K)$ has property Q or else almost every graph fails to have property Q. In a sort of *Zero-One Law*, the transition from a property

[5] We do not refer explicitly to network dynamical evolution at this stage. Indeed, here we speak about growth of random graphs only in terms of adding links to the nodes in *V* according to the given probability distribution P_K.

[6] As well as the number of edges K. Therefore we will consder throughout the following discussion that also K is function of N: $K = K(N)$

being very unlikely to it very likely is usually very swift.

To formalize the *Zero-One Law* and modelling the RG evolution by RG model, we need to consider a sequence of probability spaces $\mathcal{G}_{N,\theta}$ with $N \in \{1, 2, \ldots\}$ and θ is *K* or *p* (it depends to which sources we are referring). For each natural number *n* there will be a probability space consisting of graphs with exactly *n* nodes. In order to define if the property Q follows the *Zero-One Law*, we shall study the behavoiur of the property Q in these spaces when $n \to \infty$. Therefore, we shall say that a *typical element* of our space has property Q if the probability $P_n(Q)$ of random graphs on *n* nodes with this property tends to 1 as $N \to \infty$. When $lim_{N\to\infty} P_N(Q) = 1$ holds we say that property Q occurs almost surely (very suddenly) or, equivalently, that almost all graphs in the probability spaces $\mathcal{G}_{N,\theta}$ have property Q. When $lim_{N\to\infty} P_N(Q) = 0$ we say that property Q holds almost never, or, equivalently, that almost no graphs have property Q. Definetly in RG models, the aim is to determine at what stage of the evolution a particular property of the graph is likely to arise.

5.1.2 The Erdos and Renyi random graph evolution: the threshold functions

During the following discussion, unless otherwise stated, we will deal with the probability space $\mathcal{G}_{n,p}$, where *p* is function of *n*. As *p* goes from zero to one the random graph $\mathcal{G}_{n,p}$ evolves from empty to full. For *monotone properties* Q the value $Pr[G(n,p);Q]$, as a function of *p*, increases from zero to one [7].

As we already mentioned, Erdos and Renyi discovered that for many natural properties Q there was a narrow range in which $Pr[G(n,p);Q]$ moved from near zero to near one [32]. But this range is generally not a constant *p* but it can be described as a natural function $p = p(n)$ [83]. They called that function $p(n)$ a *threshold function* for the property Q. As an example, let Q be again the event "containing a triangle" and let $Pr[Q]$ the probability of the event Q in the graphs of the RGM probability space.

[7] A property can be monotone increasing or monotone decreasing. In [16] is proved that a monotone increasing property is the more likely to occur the more edges we have or are likely to have in an RGM process. A typical example of a monotone increasing property is the number of triangles that typically increases whenever the size of a graph increases.

Given a graph with $N = n$ nodes, there are $\binom{n}{3}$ potential triangles and each of them has probability p^3 of being a triangle; in this way, the expected number of triangles is $\binom{n}{3}p^3$. When $p(n) = n \prec\prec 1/n$ ($\prec\prec$ stands for slightly less than) (for example $p = n^{-1.01}$) the expected number of triangles is going to zero and so $Pr[Q]$ is going to zero. When $p(n) = n \succ\succ 1/n$ (for example, $p = n^{-.99}$) the expected number of triangles is going to infinity. Erdos and Renyi called $p = 1/n$ a *threshold function* for property Q. In **??** generally, we find the following definition. A function $p(n)$ is a threshold function for the property Q if it holds both this conditions:

- when $p'(n) \prec\prec p(n)$ then $Pr[G(n, p'(n)); Q] \to 0$
- when $p'(n) \succ\succ p(n)$ then $Pr[G(n, p'(n)); Q] \to 1$

Equivalently Bollobas [16] define a threshold property as: given a monotone increasing property Q a function $p(n)$ is said to be a threshold function for Q if both the statements hold:

- $p'(n)/p(n) \to 0$ implies that almost no $G(n, p'(n))$ has Q.
- $p'(n)/p(n) \to \infty$ implies that almost every $G(n, p'(n))$ has Q.

A priori there is no reason why a property should have a threshold function though with the natural definition, every monotone property of subsets of a set has a threshold function. Friedgut [38] showed necessary and sufficient conditions for the existence of sharp thresholds for monotone graph properties. This concept is closely related to the one of phase transition in statistical mechanics because the threshold functions describe how it is realized the passage from a totally disordered state (in the beginning of graph evolution) to an ordered one, in which several attributes appear in the network.

In according to Spencer [83] it is useful to think of $mathcalG(n, p)$ and also $mathcalG(n, p)$ as evolving from empty to full as $p = p(n)$ or, respectively $K = K(n)$, evolves through ever increasing functions of *n*. Indeed the most well-known property and respective threshold function is related to evolution of the connectedness of the random graph. In particular, considering the model $\mathcal{G}_{N,K}$ with $K = m$, when the number of edges increses (and N is keep fixed), every random graphs rapidly pass from small isolated clusters (at beginning of their growth), to a state dominated by one large

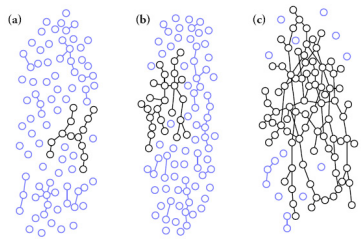


Figure 5.1: Three realizations of the random graph model of the type $\mathcal{G}_{n,K}$ with $K = m$, where $n = 100$. (a) is in the subcritical phase ($m = 40$); (b) is the evolution at the threshold function ($m = n/2 = 50$); (c) is in the supercritical phase ($m = 100$) where a giant component has alraedy formed. In the pictures the largest component is marked black, smaller components are marked grey (source: P. Holme, 2004).

connected component: the so-called *giant component* (see fig. 5.1). Mathematically, there are three regions with different behaviours: i) the subcritical phase: in this region occurring when K is of the order $N/2 - s$ where $N^{2/3} << s << N$, the expected value of the size of the largest component K_L is $O(logN)$ and the connected components are unlikely to contain more than one cycle; ii) the threshold phase: for $K = O(N/2) \pm s$, with $0 << s << N^{2/3}$, the expected value of the largest component size K_L is $O(N^{2/3})$ and graph has higher probability of contain more than one component with more than one cycle; iii) the supercritical phase: Here the expectation value for K_L is is $O(N)$, i.e. it is very likely that almost all vertices will be connected into a single giant component. Except the giant component, the graph looks like in the subcritical phase, with small components of one or no cycles.

So as above states, when $K'(n) \succ\succ K(n)$ then $Pr[G(n, p'(n)); Q] \to 1$ as $K(N)$ is overpassed. In our case Q is the property of connectedness, i.e. taken an arbitrary node the probability of being connected must be almost one. In the limit of large N $N \to \infty$, the passage of disconnected to largely conbnected occurs precisely at $K(N) = N/2$. Therefore, we can write: $K'(n) \succ\succ N/2$ then $Pr[G(n, p'(n)); (i, j) \in E, almostalways] \to 1$.

Real-world networks have very often one component much larger than any other, and a few disconnected vertices. In such cases, even if one does not know how K_L scales with N, one commonly speaks of the largest connected component as the *.giant component*. For this reason, when random graphs are used as models for real-world

networks, one use to sample the supercritical phase [49].

There exist other important threshold function governing the evolution of the characteristics of random graphs. Here in the following we resume the most important:

- At $p = n^{-2}$ edges appear, then $p = n^{-2}$ is the threshold function for the property of being nonempty;
- At $p = n^{-3/2}$ edges with a common vertex appear;
- At $p = n^{-1-1/k}$ (with k arbitrary but fixed) all trees or order $k + 1$ appear;
- At $p = n^{-1}$ triangles appear, and cycles of every fixed size k as well;
- At $p = \frac{lnn}{n}$ the graph becomes connected.

Barabasi stated in [6], that the discovery of phase transitions in graphs is fundamental because it is for this reason that graph theory tools could be applied outside of combinatorial and pure discrete mathematic. Indeed, threshold functions study represents the most important element of the whole random graphs theory. Although this order is not the one revealed in the reality, this discovery has been the basis for the modern modelisation of the emergence of non casual behaviour in real networks. In particular, random graphs represent the pure randomness and therefore the more is different it is different, the more a real network is in an ordered state. Basically real network models have been mainly devoted to answer the question about the passage from a total randomness to the ordered configurations states.

RGM then represent the basic models on which we will define phase transitions and on which we have understood the functioning of network's behavior. However we cannot use RGM as the only model for real networks. The problem with using random graphs as model is that, when we consider most of the indices discussed in Chapter 4, we see that the structural differences between RGM and real-world networks are dramatically high.

As we stated in Chapter 4, the structures represent the concepts we use for understanding the behaviour of the systems which exist on networks. In fact, the random graph models are often used as a zero-gauge against which the structural biases are defined [49]. Even objectively speaking, one can say that random graphs in many respects lack structure: an example is that its assortative mixing coefficient (see sect.

4.1.2) tend to zero in the $N \to \infty$ limit [**?**]. Perhaps, the major structural limitation that have random graph as real network model concerns its degree probablity distribution. We will discuss about this point in the next section.

5.1.3 Topology of random graphs: the Poisson degree distribution

Erdös and Rényi studied the distribution of the minimum and maximum degree in a random graph [31], while the full degree distribution was derived later by Bollobás [14]. The probability that a node *i* has $d = d_i$ edges follows the binomial distribution:

$$P(d_i = d) = \binom{N-1}{d} p^d (1-p)^{N-1-d} \tag{5.0}$$

where p^d is the probability for the existence of *d* edges, $(1-p)^{N-1-k}$ is the probability of the absence of the remaining $N-1-k$ edges and $\binom{N-1}{d}$ is the number of the ways of selecting the end points of the *d* edges. Since all the nodes in a random graph are statistically equivalent, each of them has the same distribution, and the probability that a node chosen uniformly at random has degree *d* has the same form as $P(d_i = d)$ [12].

Let's prove the following proposition.

PROPOSITION 2: *In the limit of the large nodeset cardinality, the degree distribution of a random graph in $\mathcal{G}_{n,p}$ follows a Poisson distribution.*

Proof. Let *i* be a node of a randon graph $G_{n,p}$ in $\mathcal{G}_{n,p}$ and let *p* the probability of observing an edge between *i* and the remaining $n-1$ nodes. From the Eq. 5.1.3 the mean degree $z = p(n-1)$. Therefore $p = \frac{z}{n-1}$ and the probability $P(d_i = d)$ in Eq. 5.1.3 becomes:

$$P(d_i = d) = \binom{n-1}{d} \left(\tfrac{z}{n-1}\right)^d \left(1 - \tfrac{z}{n-1}\right)^{n-1-d} =$$
$$= \binom{n-1}{d} \left(\tfrac{z}{n-1}\right)^d \left(1 - \tfrac{z}{n-1}\right)^{n-1} \left(\tfrac{n-1-z}{n-1}\right)^{-d} =$$
$$= \binom{n-1}{d} \left(\tfrac{z}{n-1} / \tfrac{n-1-z}{n-1}\right)^d \left(1 - \tfrac{z}{n-1}\right)^{n-1} =$$
$$= \binom{n-1}{d} \left(\tfrac{z}{n-1-z}\right)^d \left(1 - \tfrac{z}{n-1}\right)^{n-1} =$$

Typically, if we take the limit of large *n* holding z constant, the $P(d_i = d)$ clearly

has a Poisson distribution [41]. Hence, for $n \to \infty$ we will have that $P(d_i = d)$ converges to a Poisson with parameter z:

$$P(d_i = d) = \binom{n-1}{d} \left(\frac{z}{n-1-z}\right)^d \left(1 - \frac{z}{n-1}\right)^{n-1} \cong \frac{z^d e^{-z}}{d!} \tag{5.0}$$

□

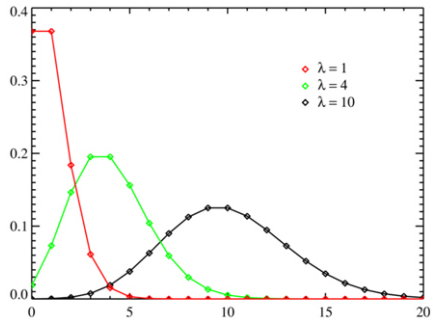


Figure 5.2: Poisson distrubution for three values of the parameter $z = \lambda$. On the orizontal axis is reported the values of the parameter.

This convergence criterion is the reason why the random graphs models are also called *Poisson random graphs*. Intuitively this result allows us to affirm that the random graph's Poisson degree distribution is quite unlike to be an highly skewed distributions of the node degrees (see fig. 5.2). In particular, it means that in random graphs of large order the nodes with mean degree are likely to be the majority, on the other hand we observe few nodes with an extreme number of connections (i.e. vertices with high or low degree are rare). Random graphs are then random in the attachment behavior among vertices: i.e., they specify a model in which agents forming a network without any preferences at all. Thus random graphs are useful as reference points against which other self-organizing structures can be compared. This is the sense of the models that we will analyze afterward.

5.1.4 Other parameter distributions in RGM

The random graphs defined in $\mathcal{G}_{n,p}$, despite the previous discussion, do not fail to have one of the principal features of real-world networks: the small-world property (see section 4.1.8. Indeed, the mean number of nodes having distance exactly h edges away from their (h-)neighbors in a random graph is z^d. Therefore the value of d needed to encompass the whole network is h and then $z^h \cong n$. It is easy to show that a typical distance through the random network is $\bar{g} = \frac{logn}{logz}$, which satisfies the definition of the small-world property given in 4.1.8 on page 65.

However, as we already saw for degree distribution, in almost all the other characteristics, RGM fail to have those typical properties of networks in the real world. RGM have a low clustering coefficient Γ: the probability of connection of two vertices is p regardless of whether they have a common neighbor, and hence $\Gamma = p$, and it tends to zero as $n - 1$ in the limit of large order [**?**]. Moreover, RGM have no correlation between degrees of adjacent vertices, and no community structure. In other words, as we suspect in sect. 5.1.3, in RGM is postulated that there is not a choice in making connection between nodes, which it represents a great limitation in modelling real world.

In short RGM are very useful in knowing how the graph characteristics change upon a modification of the network size (and order) but they are not a precise representation of real phenomena.

5.2 Generalized random graphs: static models for real networks

In sections 5.1.3 and 5.1.4 we saw the limitations of RGM as model for real networks. However, it is possible to extended RGM in several ways in order to make some of its properties more realistic. A classical example is the Exponential random graph model which is widely used in modellin social networks 6.3 on page 104.

Except the exponential version of the RGM that is rather complicated both in model

specification than in estimation parameters, there are very easy way to adjust random graph behavior: the simplest is to adapt some structural characteristic. In particular, for the discussion above, it is straightforward to conceive of modify the random graph degree distribution. In other terms we can implement a *non-Poisson degree distribution*. Random graphs with an arbitrary degree distribution $p(d)$ have been largely studied.

Bender and Canfield [30] introduced a model that allows to sample graphs with a given degree sequence D (see section 4.1.2). In the Bender-Canfield model, called *configuration model*, D is chosen in such a way that the proportion of nodes with degree d tends, for large N, to the desired degree distribution $p(d)$. The model is based on the ensemble $\mathcal{G}_{N,D}$ of all graphs with N nodes and a given degree sequence D, and each graph in the ensemble is considered with equal probability. The simplicity of this model makes it a good playground for analytical approaches. In particular the configuration model is generalisation of random graphs and consists in sample graphs of a given degree sequence: start with N disconnected vertices and assign each vertex a number of "half-edges" or "spokes" corresponding to the degree sequence. It is easy to note that pairs of half-edges form edges. In particular these half edges are randomly associated together (see fig. 5.3). The important feature of this method is that it samples the ensemble of isomorphically distinct multigraphs with a given degree sequence uniformly [**?**].

The problem with the configuration model is that, as formulated above, it can generate loops and multiple edges, i.e. it produces multigraphs and not graphs. This happens very seldom though, so for sparse and moderately sized graphs one can iterate the algorithm until a graph without loops or multiple edges is obtained [**?**].

Newman et al. [**?**] have proposed a slightly different method to generate graphs with a given degree distribution, in which the degree of nodes are independent identically distributed random integers drawn from a given $p(d)$. In this way, not a single degree sequence, but an ensemble of degree sequences is considered, and the statistical properties of the ensemble, as component sizes and number of nodes at a distance m from a given node, can be calculated by the use of a powerful formalism based on probability generating functions [96].

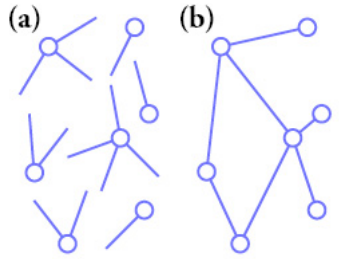


Figure 5.3: A configuration model obtained on the degree sequence $\{3, 1, 1, 4, 2, 1, 2\}$. (a) Here there are the vertices and their half-edges based on the assigned degree sequence; (b) once the half-edges are randomly attached together this is the resulting graph.(source: Holme, 2004)

One of the most important generalised random graph models is the exponential random graph family. Given its central role in estimation of social network parameters we will discuss about it in section 6.3. The important point here is that this family of models represent a static models that generalize the simple random graphs. It is very useful for the discussion we will make in section 6.3, to think about the fact that this kind of models arise from a network adaptation of the traditional models of statistical mechanics. In traditional statistical mechanics the model purpose is to interpolate between order and disorder through the (real or abstract) inverse temperature β: a more ordered configuration is associated with a lower energy ε; a configuration of energy ε occurs with a probability proportional to $exp(-\beta\ \varepsilon).Thus, the lower the temperature (the higher the \beta)$ the more likely it is that the system is in a low energy configuration. This framework has been utilised by physicists to interpolate between more random and more ordered networks ([**?**], [51]). However, the researchers that propose models that follows a similar logical pattern [49] explicitly for networks have been statisticians (see for example [66], [84] and [36]).

5.3 Random Walks on graph

The Random Walk (RW) is not a model in the sense we used along this chapter, however it is a procedure based on assigning a probability of passing through nodes to

in order to describe the sequence of nodes visited by a Random Walker in a discrete point path. We apply this procedure in order to specify some useful characteristic of the graph structure that allows to describe different models for networks (see sect. 7.1).

In RW defined on graphs the following Markov chains are considered: the state space is the set of vertices of a connected graph, and for each vertex the transition is always to an adjacenlt vertex, such that each of the adjacent vertices has the same probability. In RW are specified some interesting node quantity: expectation of recurrence times, of first-passage times, and of symmetrized first-passage times (called commuting times).

The visisted sequence is called Random Walk [33]. The random variable $s(t)$ identifies the state of the Markov chain on the graph at the time *t* that is the "position" of the random walker in the network. If the random walker is in the state (i.e. the node) *i* at the time t we have: $s(t) = i$. In particular a RW on a network is a first-order reversible Markov chain on the nodes in which the transition probability from node *i* to an adjacent node *j* (the step of random walker), i.e. the probability to pass from a state (node) $s(t) = i$ to the following state (an adjacent node *j*) $s(t+1) = j$ is defined as:

$$pr_{ij} = pr[s(t+1) = j | s(t) = i] \tag{5.0}$$

that is the conditional probability to being in the state *j* given that the random walker was in the state *i*. From the definition of node degree (see set. 4.1.2) we will rewrite Eq. 5.3 from node *i* to node *j* as:

$$pr_{ij} = pr[s(t+1) = j | s(t) = i] = \frac{a_{ij}}{d_i} \tag{5.0}$$

where: a_{ij} is the(i,j)-th element of the adjacency matrix $\mathbf{A}$ of the graph and d_i is the degree of the i-th node i.e. the row sum corresponding to the i-th row of $\mathbf{A}$. Then the matrix $\mathbf{P}$, represents the transition matrix whose entries are the transition probabilities $pr_{ij} = pr[s(t+1) = j | s(t) = i]$. In particular, $\mathbf{P}$ can be computed as:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A} \tag{5.0}$$

where: D is the diagonal matrix of the degree of the nodes of the graph.

Therefore the evolution of the Markov chain is characterized by:

$$\{x(0) = x_0; x(t+1) = \mathbf{P}^t x(t)\} \tag{5.0}$$

On the Random walk on a graph it is possible to associate several transition probability related quantitites. These quantities concern the passage of the hypothetical random walker through the nodes of the network. We will discuss about them in chapter 7. Morevoer there are very interesting connections between them and the matrices we will discuss in chapter 3. For our purposes these are the only elements we need to define node distances and other important elements on networks. However the RW models have a number of useful properties, not directly connected with the dissertation topics, that are fully described in [33], [53] and [60].

5.4 Small world networks

We faced several RGM limitations in order to modelling real networks (see sect. 5.1.3 and 5.1.4). We also saw that in generalized random graphs (esct. 5.2), some authors tried to obtain more reliable models by applying some specific "structural correction" to the RGM. However, these latter models are sometimes very untractable (i.e., exponential random graph model) or they do not generate graphs (i.e. configuration models).

In the following, we analyze easier models for real networks. In particular we will focus on the most important models that have tried to overpass RGM structural limitations: the class of *Small-World networks models* (SWM). The progenitor of such models is the *Watts and Strogatz model* (WS) [94], [93], [95]. We now present its principal topological features.

Though also with exponential random graphs one can obtain networks with logarithmically increasing path-length and an arbitrary clustering, a simpler method to get the same result is the WS model. Therefore the WS model is a method to construct graphs, denoted as $\mathcal{G}^{WS}(N, K)$ having both the small-world property and a high clustering coefficient [94]. The model is based on a rewiring procedure of the edges implemented

with a probability p. The starting point is a N nodes regular lattice, in which each node is symmetrically connected to its $2m$ nearest neighbors for a total of $K = mN$ edges. Then, for every node, each link connected to a clockwise neighbor is rewired to a randomly chosen node with a probability p, and preserved with a probability $1 - p$. Therefore, for $p = 0$ we obtain the starting point regular lattice, whereas for $p = 1$ a random graph with the constraint that each node has a minimum degree $d_{min} = m$ is build up. For $0 < p < 1$ the model produces networks with the small-world property and, above all, a non-trivial clustering coefficient (see fig 5.4 to appreciate the generating process). The rewiring process allows the small-world model to interpolate between a regular lattice and something which is similar, though not identical, to a random graph. In a more fashionable way we can affirm that small world networks lie between the extreme regularity and the total randomness. Moreover in these kind of networks arise several properties on which are based many real systems [94], [95].

Since its appearance in the seminal paper of Watts and Strogatz, this model has stimulated an intense activity aimed at understanding the networks' properties as a function of the rewiring probability p and the network order N. The small-world property mostly derived from the immediate decreasing in the average shortest path length (viewed as function of the rewiring probability) $\bar{g}(p)$ as soon as p is slightly larger than zero. This happens because the rewiring of links creates many shortcuts that connects otherwise distant nodes. The effect of the rewiring is highly nonlinear on $\bar{g}$, and not only affects the nearest neighbors structure, but it also opens new shortest paths to the next-nearest neighbors and so on. Conversely, an edge redirected from a clustered neighborhood to another node has, at most, a linear effect on the clustering coefficient Γ. The transition from a linear to a logarithmic behavior in $\bar{g}(p)$ is faster than the one associated with the clustering coefficient $\Gamma(p)$. This leads to the appearance of a region of small (but non-zero) values of p, where one has both small path lengths and high clustering.

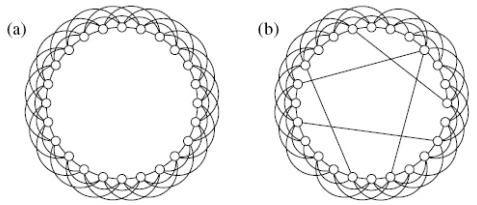


Figure 5.4: (a) The one-dimensional regular lattice from which the Watts and Strogatz model starts in order to generate small world networks; (b) the process of generation of small world networks consists in randomly 'unplug' some regular lattice connection with a probability p (and $1 - p$ is the probability that links remain in the original state) and reconnecting it toward a new nodes, also chosen uniformly at random (source: Newman, 2003).

5.4.1 Topology of small world networks

The degree distribution of the small-world model does not look like the one in most real-world networks. Moreover, because of the rewiring procedure its functional form is somewhat complicated.

The clustering coefficient Γ of the small world graphs is equal to:

$$\Gamma^{WS} = \frac{3(d-1)}{2(2d-1)(1-p)^3} \tag{5.0}$$

Obviously Γ depends on the probability p.

Apart form the clustering coefficient, the attention in the WS model has been focused on the average geodesic path length $\bar{g}$. It is easy to note that in the limit of $p \to 0$ the model could be called "large world" because the average path length tends to $\bar{g} = N/4d$, as discussed above. Small-world behavior, by contrast, is typically characterized by logarithmic scaling $\bar{g} \approx logN$ (see section 4.1.8), which we see for large p, where the model becomes like a random graph. In between these two limits there is presumably some sort of crossover from large-world to small-world behavior.

To describe the behaviour of the average geodesic in WS model we now simulate an example of how small world networks are obtained starting from a k-nearest-neighbor ring network as could be the regular graph G in fig. 2.1. By the adjacency matrix $\mathbf{A}$ of G, in table 2.1, we see that our graph $G(V, E)$ has 9 nodes and is a k-nearest neighbor with $k = 2$, because in it every node is reachable from any other node in maximum

$k = 2$ steps. By using this definitions, we can generate a general *k*-nearest-neighbor lattice with arbitrary N and k. We use here a variation of the WS model generating procedure that is based on the use of the shortcuts in the regular lattice rather than on the rewiring procedure above described [**?**]. Now we superimpose random shortcuts on the network *G*: in other words we add nonzeros entries to the adjacency matrix at random, according to the following process. We consider N flips of a biased coin that outcomes "head" with probability p. If th *i*-th flip yield a "head" then we choose a column $1 \leq j \leq N$ uniformly and set $a_{ij} = a_{ji} = 1$. Operatively, we add a link, that is a shortcut, from the *i*-th node to the *j*-th randomly chosen node. On average the total number of generated shortcuts is equal to Np.

Generate shortcuts in this way make it very likely to obtain a graph with low average pathlength. Our aim is now to investigate for fixed N, or in other words for a family of random graph of the type $\mathcal{G}_{N,p}$, how fast is the process of average pathlength decay as the number of shortcuts increases.

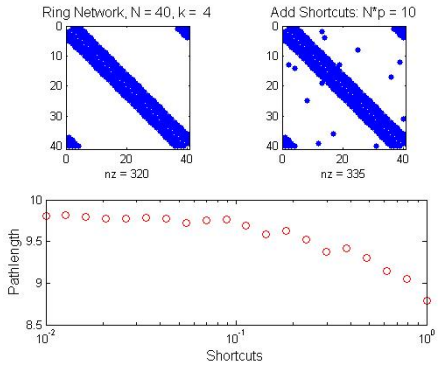


Figure 5.5: In the upper part is showed the process of adding shortcuts to the regular graph. In particular, it is depicted the adjacency matrices of a regular k-nearest-nighbour on 40 nodes (on the left part) and of the small world network, on the same number of nodes, when randomly assigned shortcuts are added (on the rigth part). In the lower part is represented the average shortpath length decay of a k-nearest-neighbor, on 150 nodes and with k=4, after the addition of 10^c shortcuts. This part of the figure is plotted on semilog plane.

After the simulation we can see (fig. 5.5) that the pathlength drops significantly when an average of $O(1)$ shortcuts are added to the starting regular $O(N)$ network. Therefore small world networks are characterized by an high clustering coefficient

and also for the presence of few ”strategic” shortcuts that dramatically drop down the average shortpath length in the system.

5.5 A particular case of Small World Networks: the Scale Free Networks

Has we already manetioned in section 4.1.2, it has been proved that many real-world networks have a power-law degree distribution, i.e. the probability density function $p(d)$ has the form:

$$p(d) = ad^{-\gamma} \tag{5.0}$$

where a and γ are constants. This kind of distribution is very far from what would be expected from a purely random network formation process. Indeed, we demonstrated that degree distribution in RGM follows a Poisson law. Th crucial point is that power-laws have a special role in statistical mechanics: the traditional models of statistical physics usually exhibit a phase transition between ordered and random states. Exactly at the transition the system has ”self-similar” or ”scale-free” properties. The process of rescaling on a real function f is mathematically treated by tuning a scale parameter $a \in (0, \infty)$; so f is scale-free, which intuitively means that has the same form for all scales, if it follows:

$$f(ax) = bf(x), \forall x \in \Re \tag{5.0}$$

where b is a constant. The only solution of the Eq. 5.5 is a power-law as Eq. 5.5. Networks with a power-law degree distribution are, for this reason, often called *scale-free networks*. Given some class of graphs with power-law degree distribution it is, however, not necessarily true that the network is scale-free [49].

Networks with power-law degree distributions have been the focus of a great deal of attention in the literature. They are sometimes referred to as *scale-free networks*, although it is only their degree distributions that are scale-free. It is possible to require

scales being present in other network properties. The earliest published example of a scale-free network is probably Price's network of citations between scientific papers [27]. He computed a value of $\alpha \in [2.5, 3]$ for the exponent of his network.

More recently, power law degree distributions have been observed in a high number of other networks, including other citation networks, the World Wide Web, the Internet, metabolic networks, telephone call graphs, and the network of human sexual contacts. Other common functional forms for the degree distribution are exponentials, such as those seen in the power grid, and railway networks, and power laws with exponential cutoffs, such as those seen in the network of movie actors and some collaboration networks.

As with all systems characterized by a power law distribution, the most notable characteristic in a scale-free network is the relative commonness of nodes with a degree higher than average. The highest-degree nodes are often called *hubs*, and are thought to serve specific purposes in their networks, although this depends greatly on the domain.

The maximum degree d_{max} of a vertex in a network will in general depend on the size of the network. For some calculations on networks the value of this maximum degree matters. In work on scale-free networks, Aiello, Chung, and Lu assumed that the maximum degree was approximately the value above which there is less than one vertex of that degree in the graph on average, i.e., the point where $np(d) = 1$ [90]. This means, for instance, that $d_{max} \approx n^{1/\alpha}$ for the power-law degree distribution instead $p(d) \approx d^{-\alpha}$. This assumption, however, can give misleading results; in many cases there will be vertices in the network with significantly higher degree than this. Given a particular degree distribution (and assuming all degrees to be sampled independently from it, which may not be true for networks in the real world), the probability of there being exactly *m* vertices of degree *d* and no vertices of higher degree is $\binom{n}{m} p_d^m (1 - P(d))^{n-m}$, where $P(d)$ is the cumulative probability distribution. Without using differntial calculus a simple rule of thumb that leads to the same result is that the maximum degree is roughly the value of *d* that solves $nP(d) = 1$. Note, however, that, as shown by Dorogovtsev and Samukhin, the fluctuations in the tail of the degree distribution are very large for the power-law case [74].

The presence of these kind of hghly conncted nodes, i.e. the power law degree distribution, dramatically influences the network topology. It turns out that the major hubs are closely followed by the little smaller ones, and these latter are followed by other nodes with an even smaller degree and so on. Since the vast majority of nodes are those with small degree, even if dertructive event occurs in the system, it is very likely that the network will not lose its connectedness, which is guaranteed by the remaining hubs.

Another important characteristic of scale-free networks is the clustering coefficient Γ distribution, which decreases as the node degree increases. The distribution of Γ also follows a power law. That means that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs.

At present, the more specific characteristics of scale-free networks can only be discussed in either the context of the generative mechanism used to create them, or the context of a particular real-world network thought to be scale-free. As with most disordered networks, such as the small world network model, the average distance between two vertices in the network is very small relative to a highly ordered network such as a lattice. The clustering coefficient of scale-free networks can vary significantly depending on other topological details, and there are now generative mechanisms that allow one to create such networks that have a high density of triangles.

Although many real-world networks are thought to be scale-free, the evidence remains inconclusive, primarily because the generative mechanisms proposed have not been rigorously validated against the real-world data. As such, it is too early to rule out alternative hypotheses.

5.5.1 Static Scale Free Networks

The large amount of works on the characterization of the topological properties of real networks has motivated the need to construct graphs with power law degree distributions. Graphs with a power law degree distribution can be simply obtained as a special case of the random graphs with a given degree distribution discussed in section 5.2. We denote such graphs as static scale-free to distinguish them from models of

evolving graphs that will be discussed in section 5.5.2. Aiello et al. have studied a model that involves only two parameters, α and γ, and that assigns uniform probability to all random graphs having a number of nodes with degree *d* given by $N(k) = e^{\alpha} d^{-\gamma}$. We denote this model as $G_{\alpha,\gamma}$ since the total number of nodes *N* and edges *K* are fully determined, once the parameters α and γ are given. Notice that α is the logarithm of the number of nodes with $d = 1$, while $e^{\alpha/\gamma}$ is the maximum degree of the graph. Newman has calculated the clustering coefficient for power-law degree distributions finding that Γ tends to zero in large graphs with $\gamma > 7/3$, while for $\gamma < 7/3$, Γ increases with the system order, due to the fact that there can be, in average, more than one edge between two nodes sharing a common neighbor. Chung and Lu have studied the average distance in random graphs with a power-law degree distribution, proving that it scales as $logN$ if $\gamma > 3$. More recently, Chung et al. have shown that the spectrum of the adjacency matrix of random power lawgraphs follows a power law distribution, while the eigenvalues of the normalized Laplacian (see section 3.2.1) follow the semicircle law. Several other recipes to construct static scale-free networks, based on assuming that a node has some intrinsic properties, have been proposed in physics.

5.5.2 Dynamical Scale Free Networks: the preferential attachment model

Complex networks are in general systems in evolution, with new nodes/edges that get formed and old ones that get removed or destroyed. The currently accepted mechanism finds its roots in an old idea of de Solla Price, based on the so-called preferential attachment, which means that a newly formed node i builds an edge with a preexisting node with a probability that is proportional to the degree of the latter node [27]. Networks constructed in this way have a degree distribution with a power law tail, as observed in real networks. This simple rule, however, makes implicitely the strong assumption that each node is at any time informed about the degree of all other nodes, which is certainly not true, especially for gigantic systems which contain many millions of nodes, like the WWW. We rather believe that the key behind the building of

a connection between a pair of nodes lies essentially in the mutual interaction of the two nodes, almost independently of the rest of the system: two persons usually become friends because they like each other.

As we saw before, static scale-free networks are good models for all cases in which growth or aging processes do not play a dominant role in determining the structural properties of the network. Conversely, there are many examples of real networks in which the structural changes are ruled by the dynamical evolution of the system. The class of models whose primary goal is to reproduce the growth processes taking place in real networks: the rationale is that one will be able to reproduce the topological properties of the system as we see them today. We concentrate primarily on the model of network growth proposed by Barabasi and Albert in 1999 [4].

The *BarabasiAlbert model* (BA) is a model of network growth inspired to the formation of the World Wide Web and is based on two basic ingredients: growth and preferential attachment. The basic idea is that in the World Wide Web, sites with high degrees acquire new links at higher rates than low-degree nodes. More precisely, a undirected graph $G^{BA}_{N,K}$ is constructed as follows. Starting with m_0 isolated nodes, at each time step $t = 1, 2, 3, \ldots, n - m_0$ a new node *j* with $m \leq m_0$ links is added to the network. The probability that a link will connect *j* to an existing node *i* is linearly proportional to the actual degree of *i* [4]:

$$\Pi_{i \to j} = \frac{d_i}{\sum_l d_l} \tag{5.0}$$

Because every new node has *m* links, the network at time *t* will have $N = m_0 + t$ nodes and $K = mt$ links, corresponding to an average degree $\bar{d} = 2m$ for large times. As for many other scale free networks also the BA model has many similarities with a model developed by Price in 1976 to explain the power laws that the same author found, one decade earlier, in citation networks (both for the in-degree and the out-degree distributions).

In Price's model, the probability that a new published paper cites a previous one is taken to be proportional to $d_{in} + 1$, where d_{in} is the number of times that the paper has already been cited [27]. Price's model is a reformulation, in terms of network growth,

of a model developed by Simon in 1955 to explain the power laws that appear in a wide range of empirical data, as in the distribution of words in prose samples by their frequency of occurrence, or in the distributions of cities by population. The detailed discussion on the solution of Price's model can be found in the review by Newman [**?**].

In the limit $t \to \infty$, the BA model produces a degree distribution $P(d) \approx d^{-\alpha}$ with an exponent $\alpha = 3$ The average distance in the BA model is smaller than in a random graphs models with same N and K, and increases logarithmically with N. The clustering coefficient vanishes with the system order as $\Gamma \approx N^{\cdot 75}$. This is a slower decay than that observed for random graphs, $\Gamma \approx N^{-1}$, but it is still different from the behavior in small-world models, where Γ is a constant independent of *N*.

For instance, networks generated by preferential attachment typically place the high-degree vertices in the middle of the network, connecting them together to form a core, with progressively lower-degree nodes making up the regions between the core and the periphery. Many interesting results are known for this subclass of scale-free networks. For instance, the random removal of even a large fraction of vertices impacts the overall connectedness of the network very little, suggesting that such topologies could be useful for security, while targeted attacks destroys the connectedness very quickly. Other scale-free networks, which place the high-degree vertices at the periphery, do not exhibit these properties; notably, the structure of the Internet is more like this latter kind of network than the kind built by preferential attachment. Indeed, many of the results about scale-free networks have been claimed to apply to the Internet, but are disputed by Internet researchers and engineers.

The BA model generates networks with power-law degree distribution with exponent approximately -3 (see scet. 4.1.2) and a clustering (sect. 4.1.7) that slowly converges to zero. The BA model is not the only growth-model that generates power-law degree distributions; it has some features that real-world systems hardly share. For example there will be a strong correlation between age and degree, and furthermore only linear preferential attachment will give a power-law degree distribution.

CHAPTER 6

Social Network Analysis

A social network is a group of people with some pattern of contacts or interactions between them [78]. The patterns of friendships between individuals, business relationships between economic organizations, the web of sexual contacts and intermarriages between families are all examples of networks that have been studied in the recent but intenses years of activities of the so-called social network analysis (SNA). Typically, network studies in sociology have been data-oriented, involving empirical investigation of real-world networks followed often aimed at determining the centrality or influence of the various actors.

In this part we adapt the graph theory concept to the description of social interactions. Some author distinguish between graphs and the social network by saying that a social network consists of a graph and additional information on the vertices or on the lines. In the previous chapters we discussed about the part concerning the network as a complex system: in particular, complex network theory, by its graph theoretical approach, does not explain the network by its elements' behaviour, but it deals with a whole organism that evolves by means of its single components. In this chapter instead we approach the study of that components: indeed SNA, as we already stated, deals with the study of the actors involved in a network. It is an approach that is based on the individual behaviour: it is the single behaviour that generates the network and governs its evolution. The single components have the possibility of choose their own connection without considering the network as a structure but only because some individual charateristic.

Indeed almost the totality of social network models deals with the external information on the actors. Social network analysts often use these additional information to explain network formation.

This is perhaps the principal difference between social network analysis, from one hand and complex network theory form the other hand. In the latter we often disregard the additional information on the single nodes because the attention is point out toward a most structural approach that we call a ”systemic approach”. This approach is what we think it should be adopted also in SNA. Social networks configurations are very often explained as dependent by the actors' characteristics and rarely it is applied a more systemic approach. Indeed, especially for large networks of interactions the probability of ties among actors is not directly a consequence of a shared characteristics but also depends on the network topology. We mean that the probability of ties between two persons it depends not only on the common fetaure they share but also by the conformation of the network at certain time and space (i.e. by the network topology). In many situations the personal behaviour does not influence the social environment in which one is involved.

6.1 Basic analytical perspective in social network analysis

SNA in the field of social sciences have a long history in the quantitative study of real-world networks. Of particular note among the early works on the subject are the following: Moreno's networks of friendships within small groups, of which Fig. 6.1 is an example; the so-called southern women study of Davis, Gardner, and Gardner, which focused on the social circles of women in an unnamed city in the American south in 1936; the study by Elton Mayo and colleagues of social networks of factory workers in the late 1930s in Chicago; the mathematical models of Rapoport, who was one of the first theorists, perhaps the first, of systemic social network analysis in social sciences, the notable study of of patterns of sexual contacts [77].

Another important set of experiments are the famous, and already cited, ”small-

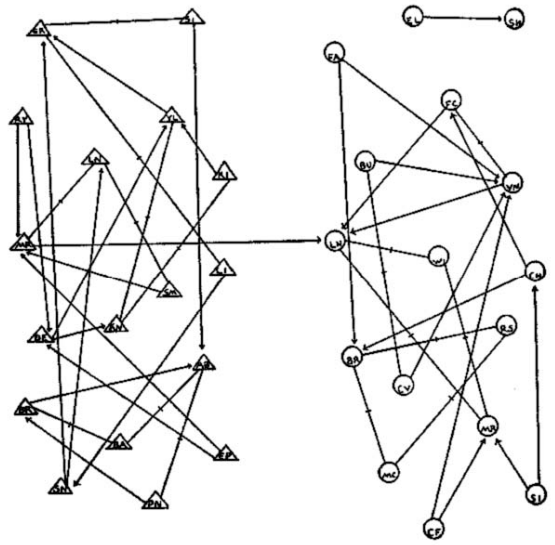


Figure 6.1: An early hand-drawn social network from 1934 representing friendships between school children. The triangles and circles represent nodes with different attributes (Reprinted with permission from ASGPP)

world" experiments of Milgram [**?**]. No actual networks were reconstructed in these experiments, but nonetheless they tell us about network structure. The experiments probed the distribution of path lengths in an acquaintance network by asking participants to pass a letter to one of their first-name acquaintances in an attempt to get it to an assigned target individual. Most of the letters in the experiment were lost, but about a quarter reached the target and passed on average through the hands of only about six people in doing so. This experiment was the origin of the popular concept of the "six degrees of separation" although that phrase did not appear in Milgram's writing, being coined some decades later by Guare.

Traditional social network studies often suffer from problems of inaccuracy, subjectivity, and small sample size. With the exception of a few ingenious indirect studies such as Milgram's, data collection is usually carried out by querying participants directly using questionnaires or interviews. Such methods are labor-intensive and therefore limit the nuber of actors can be observed. Survey data are, moreover, influenced by

subjective biases on the part of respondents; how one respondent defines a friend, for example, could be quite different from how another does. Although much effort is put into eliminating possible sources of inconsistency, it is generally accepted that there are large and essentially uncontrolled errors in most of these studies. A review of the issues has been given by Marsden [5].

The starting point of SNA is a set V of n actors and a set E of m links among them. Then E is a set containing couples of elements of V. The mathematical object $G(V, E)$ composed of these sets is the network (or a graph). This data are represented in a sociomatrix which is the analogue of the adjacency matrix in graph theory.

The main purpose of SNA is to analyze the pattern of recurrence structure in an actor's network by means of the interpretation of the so-called local structures and by using, most of the times a fequency approach of the elementary interaction structures. The basic elementary interaction structures are the *dyads* and the *tryads* [64], [65], [85]. Substantially almost every work in SNA is an analysis of the configuration of this structure in the whole social network. This reductionism allows to decompose the social system in singles parts and the frequency of these configuration define the property of the whole system. We consider in this work this reductionism as a convenient analytical choice but not as the only way to interpret network's properties. We will try to explore more flexible ways of analysis by using advanced tools.

However, several attempts has been directed to overpass this reductionism by analyzing more general substructures in the sociogram. A more global or macro paradigm it is represented by the analysis of the blockmodels [75], [61].

A systemic point of view, as we suggested, has been applied by the notable work of Wasserman [86] in which there is a specification of a stochastic process on a social network in order to modelling a network structure evolution. We will make something similar to the work of Wasserman when we will apply a stochastic mechanism on graph (namely RW) in order to interpreting actors' distances (cfr. sect.).

6.2 SNA as local structures analysis

Historically the statistical network analyses in SNA have been referred to as "local methods". Indeed, these methods look at the subgraphs rather than at the entire collection of the actors involved in the network. Also in the recent development of the statistical approach to the SNA the attention is focus toward the local structures in a network. We already sugget that this approach depend greatly by the sociogical method of study social interactions. Briefly, a *local structure* is usually defined to be the regularities in a social system of actors and relations that can be considered as a subgraph rather than as the global graph (or directed graph) itself [91].

As we above stated, in the majority of the social network statistical models, the level of analysis is local: dyads (subgraphs composed of pairs of actors and all links between them) and triads (subgraphs of three actors and all ties among them). These local structures furnish a local view of the whole network system .

An important concept related to the locality of a network is the one of *structural equivalence*. Structural equivalence measures are designed to quantify how similar the positions of two actors in the network are. In the binary sense, two actors *i* and *j* are structurally equivalent if they have the ties to the same actors in $V' = V - \{v, w\}$ [**?**]. Briefly, structural equivalence is a mathemaical porperty of single actors (or subsets of actors): in particular, two actors are structurally eqivalent if they have identical ties to and from all other actors in the network. Extensions of this concept have been proposed: in particular the possibility of create a continuous measure S_t of structural equivalence by measuring the Euclidean distance between the in-neighbourhoods and out-neighbourhoods [19] is closely related to the methodological contribution presented in chapter 7.

It is worth to note that, in dealing with local structure, it is important to consider the direction of the ties. Indeed the states of local structures are dramatically dependent on the direction of the ties. This is quite obvious if we think about dyads. In simple graph a dyad has only two states: existencs or nullity. In digraph the possible dyadic states increase to four. Therefore, often in SNA are considered digraphs rather than undirected networks.

6.2.1 Dyads

A dyad is the most basic element of connectivity in a social network. Given that it is the smallest possible social group, it is used as basic element in the construction of every social network statistical method. In SNA a dyad is defined as an unordered pairs of actors and the arcs that exist between the two actors in the pair. The dyad consisting of actors (i, j) will be denoted as $D_{ij} = (e_{ij}, e_{ji})$, for $i \neq j \in V$. Thus D_{ij} is a 2-subgraph.

In undirected graphs, the number of D_{ij} is closely related to the number of edges. Indeed the density (see section 4.1.3) can be interpret as an index that measures the propensity of occurrence of dyads within a social network. The simple count of edges therefore corresponds to the number of non-null D_{ij} in a graph.

If we consider the directed case we have to count the total number of dyads (the size of the graph) but we have to distinguish between outcoming dyads and incoming dyads to respect a given actor.

Thus, given the direction of links, in general D_{ij} can have three different states: mutual (M), asymmetric (A) and null (N). Mathematically, these possible states of D_{ij} are called *isomorphism classes* because all the possibilities of existence of 2-subgraph is referred to one of these three configuarations.

The simplest dyads is the null, that is when two actors are disconnected. A dyad is mutual if both the tie from i to j and the tie from j to i are present. Asymmetric means that either i has a tie to j or j has a tie to i, but not both. These isomorphism classes are often labeled M, A, and N.

The *dyad census* gives the frequencies of these types. By using this frequency a number of indices can be calculated. For example an index of mutuality, m_{kp}, has been suggested by Katz and Powell [56]. This index focuses on the probability of a mutual choice between two actors: Prob(a chooses b and b chooses a) as the product of two probabilities: $Prob(a \text{ chooses } b) \times Prob(b \text{ chooses } a | a \text{ chooses } b)$. The second half of this product can be thought of as consisting of two parts: the $Prob(b\text{chooses } a)$ and a fraction of the a priori probability that b does not choose a [56]. This fraction is 0 if there is no tendency toward mutuality and 1 if there is a perfect tendency toward

mutuality. The index m_{kp} is interpreted as this fraction. The index is a linear function of the number of mutuals in the network.

6.2.2 Triads

The subgraphon three vertices, a 3-subgraph, is called a *triad*. The triads are complex local structure especially for the explosion of isomorphism classes that are obtained when it is considered a directed graph. Indeed, if we select three vertices and the links among them in the directed network we always get one of the 16 combinations (see fig. 6.2).

While in dyadic labelling is usually used the MAN denotation, for triads there exist two different ways of labeling them. The first consistes simply in numbering $1, \ldots, 16$, the second type of label consists of three numbers xyz, where:

- x is the number of pairs of vertices connected by bidirected edges;
- y is the number of pairs of vertices connected by a single edge;
- z is number of unconnected pairs of vertices.

When the three numbers are not enough to distinguish among triads it is used an additional letter: D for down, U for up, C for cyclic, T for transitive. As for dyads also for triads there exist the censuse and it is called *triad censuses*. It is mainly based on the counting of transitive triads rather than on all the 16 isomorphism classes.

A triad, involving nodes u, v, w is transitive, if whenever exist arcs $u \to v$ and $v \to w$, exists also $u \to w$. This transitivity is measured by the clustering coefficient intorduced in section 4.1.7. Triads that are not transitive are called intransitive. Considering the fig. 6.2: we can see that some intransitive triads are ”more intransitive”, some are ’less intransitive’, e. g. in triad 15 only one arc is missing to be transitive, in triad 11 two arcs are missing. In particular:

- triads 9, 12, 13, 16 are transitive ;
- triads 6, 7, 8, 10, 11, 14, 15 are intransitive;
- triads 1, 2, 3, 4, 5 do not contain arcs to meet of transitivity (they are vacuously transitive).

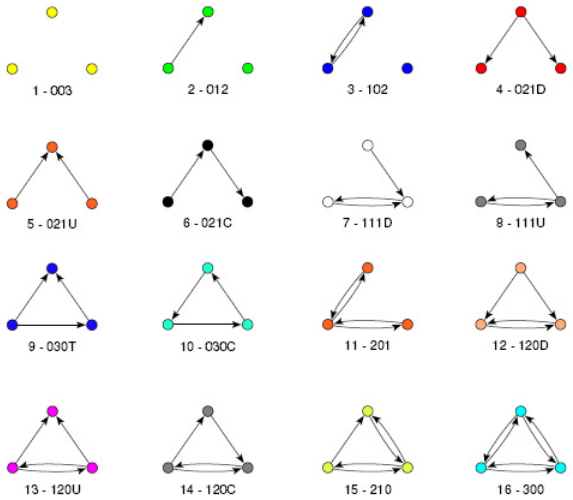


Figure 6.2: The 16 possible isomorphism classes of the triad used in the triad censuses(source: Mrvar, 2003).

6.2.3 Cliques

Subset of vertices in a network is called a clique, if every vertex from the subset is connected to all other vertices in the subset (clique is a special type of a core). A clique (in social networks) represent a subset of persons who are connected as much as possible. In general complex network theory these maximal subsets are called ”motifs” (see section 4.1.9). Searching for cliques is computationally very expensive than searching simple local structures. Therefore we will only search for cliques of size 3 or 4 at most in smaller networks. In particular a 3-clique is the triad number 16 (see fig. 6.2). Using counting of triads we can find how many 3-cliques exist in a network.

6.3 Models for social networks

Statistical analysis of network ties is both important and difficult. It is important (among other reasons) because it can allow us to distinguish pattern from random noise,

and because it enables us to assess comparatively a variety of hypotheses about the structures that underlie or generate the network data that we observe. It is difficult because the units of observations (such as the tie from *i*-th to *j*-th node and those from *i*-th to *k*-th and from *j*-th to *k*-th) are not independent.

The first attempt to modelling in social networks was in 1930's by Moreno and Jennings: In that paper they recognized that "a change in position of one individual may affect the whole structure" [**?**]. More recently, in a series of works, Holland and Leinhardt [64] [66], formulated an exponential family of models (which they called the p_1 family) that provided estimation of the probability of an observed network conditioned on the number of ties sent and received by each actor (and the network's overall density) and, simultaneously, on the two-person configurations (mutual, asymmetric, and null-choice dyads) among the actors.

Several problems with the p_1 model conducted to further works. First, the assumption of dyadic independence (the assumption, e.g., that person *i*-th person's choice of *j* is independent of *i*-th person's choice of *k*), which is not fair in many contexts. Second, because the p_1 model degrees of freedom for assessing the overall fit often depended on the number of actors in the network, and this is a violation of the usual assumptions about asymptotics in maximum likelihood estimation.

Recent work constituting fully a new breakthrough, however, has allowed the assumption of dyadic independence. These new models, as formulated in [87] and [62] are called *exponential random graph models* (ERGM) or $p*$ models. In such model it is asserted that a specific collection of simple, concrete structures (called "configurations" or "local structures") compose the network as a whole. As we focus on the beginning of this chapter this peculiarity charcaterisation of the networks is the principal deifferences between SNA and the other branches of network theory. Single networks as well as networks of multiple types of relation (such as co-worker and socializing, in Lazega and Pattison's 2001 study of 71 partners and associates in a law firm with offices in three cities [**?**]) may be modeled with this approach.

There is also an equilibrium aspect related to ERGM, in that the presence or absence of each possible tie in the network is estimated conditioning on the rest of the data. It

may be the case that, once these local structures are specified, very standard methods of estimation (e.g. "logit regression") are applicable. However, the estimates do not satisfy some of the usual assumptions. Indeed, they are pseudo-likelihood estimates rather than maximum likelihood estimates.

6.3.1 Generality on the p* class of models

The models used to describe the observed networks in SNA are, as we stated above, mainly based on the so-called $p*$ class of models, which are a case of exponential random graph models. The $p*$ models and therefore the ERGM are closely related to the random graph (RGM) described in section 5.1. In particular they are random graph with several dependency constraints and statistical analysis specification on it.

Like in RGM, also in ERGM the observed network is considered as one realization from an ensemble of possible networks having in common some structural characteristics (at very least, the same number of actors). In other words, it is a realization of an unknown stochastic process. The range of possible networks and their probability of occurrence under a determined model is represented by a probability distribution.

In general, the goal is to evaluate the unknown stochastic process that governs the observed network structure by using some parameters ϑ, which represent some structural characteristics. More important is that the structural features in question, help to shape the form of the model. For example, an assumption of a reciprocity process laeds us to propose a model where the level of reciprocity is a parameter [40].

The selected model parameters are estimated by using the observed network as a guide in the sense of maximum likelihood criterion (ML). ML for social networks is the same as in ordinary statistics. It consists in the choose the parameter values in such a way that the most probable value is the one which occurs in the observed networks. Basically, once defined a good-fit model in terms of ϑ, we evaluate the probability that the parameter of interest occurs by chance alone in the observed network. Moreover, we can define a probability distribution on the set of all graphs (with fixed structural characteristics, i.e. the same nodeset cardinality) and also drawing graphs at random from the distribution according to their assigned probabilities. We can compare the

sampled graphs to the observed network on any other charactersitic of interest. If the model is good fit for the data, then the sampled graphs will be very similar to the observed one in many different aspects. This procedure, ideally, allows to infer that modeled structural effects could explain the emergence of the network.

Following Robbins et al. [40], let consider friendship in a school classroom. The observed network is the one for which we measured the frienship interactions. There are many different networks that could be observed from for that classroom: some friendship structures may be very likely or unlikely to be observed, and the set of all possible structures with some assumption about their associated probabilities, forms a probability distribution of graphs for the ensemble. The ensemble is build in such a way all the graphs in it have some particular structure in common. We place the observed network within this distribution, under the basic constraint that the model is a good one for our data.

6.3.2 The p* class of models construction

The initial assumptions that allows to build up a $p*$ model are quite similar to the ones described for Random Graphs in sec. 5.1. As starting point we consider a stochastic structure in which at least the nodeset cardinality is fixed. Other structural properties could be fixed as well (i.e. the same number of outdegree, the edgeset cardinality, etc.) but in this way we dramatically reduce the number of possible networks in the ensemble.Basically we assume that a relational tie between the *i*-th actor and the *j*-th actor is a random variable Y_{ij} with two possible values: $Y_{ij} = 0$ if there is no connection between *i* and *j* and $Y_{ij} = 1$ if there is a tie. We denote y_{ij} the observed value for the variable Y_{ij}, with $\mathbf{Y}$ the matrix of all variables and with $\mathbf{y}$ the matrix of ties in the observed network. Briefly $\mathbf{y}$ represents the adjacency matrix of the observed network.

Differently from random graph models here we assume that there is some dependence hypothesis under the ties formation. at the beginning of the $p*$ models, namely in p_1, ties was assumed to be independent of each other. It can be showed that well-specified dependence assumptions imply a particular class of models (see the Hammersley-Clifford theorem in [8]). Each parameter corresponds to a *local structure*

in the network, i.e. to a small subset of possible network edges. These local structures represent the structural characteristics under study (e.g. the reciprocated ties or the transitivity). For instance, a single tie can be a local structure, as may be a reciprocated tie (in digraphs), a transitive triad and a two star. Parameters related to the appearance of these local structures in the observed network can be included in the model.

The exponential random graph models (ERGM) have this form:

$$Pr\left(\mathbf{Y}=\mathbf{y}\right)=\left(\frac{1}{\kappa}\right)exp\left\{\sum_{A}\eta_{A}g_{A}(\mathbf{y})\right\} \tag{6.0}$$

where the summation is made over all the local structures A, κ is a normalizing costant, η_A is the parameter related to the local structure A and $g_A(\mathbf{y}) = \prod_{y_{ij}\in A} y_{ij}$ is the *network statistic* corresponding to A. In particular, $g_A(\mathbf{y}) = 1$ if the local structure is observed in the network $\mathbf{y}$, $g_A(\mathbf{y}) = 0$ otherwise. All ERGM are of the form in Eq. 6.3.2 which describes the general probability distribution of graphs on *n* nodes.

If we allows in Eq.6.3.2 the possibility to include more general statistics than local structures A and subgraph counts then the dependence structure may not be clear. However in this case we only consider just a very simple model specification inducted by allowing only the presence of the subgraphs A.

In the Eq.6.3.2 is explicit that the probability of observing a given network $\mathbf{y}$ dependends both on the statistics in $g_A(\mathbf{y})$ in the network $\mathbf{y}$ and on the all non-zero parameters η_A for all local structures included in the model. It is worth to note that η_A is non-zero if and only if all pairs of variables in A are supposed to be *conditionally dependent*. That is why the dependency assumption is important. The presence of non-zero η_A specifies that the only local structures Athat are important are those in which possible ties in it are mutually dependent on each other. The entity of the parameter value affects the probability that the specified local structure A (the one the parameter is referred to) has to be observed in an hypotetical network evolution. Of course, the higher the parameter, the more the expectation of observing the corresponding A.

6.3.3 Dependence assumption and models

The Eq. 6.3.2, as we mentioned before, is also referred as $p*$ models, because the ERGM are a generalization of the p_1 models of Holland and Leinhardt. The most simple case of ERGM are the so-called *Bernoulli graphs* which represent an exponential generalization of the random graphs. To the class of ERGM belongs also other kinds of models as the *Markov random graphs* introduced by Frank and Strauss in order to modelling a directed network by dependency constraints. The purpose of this section is to briefly explore the major specifications of ERGM from the simpler to more analitically complicated.

Bernoulli graphs

Bernoulli random graphs are generated when we assume that ties are independent. This is the case of random graphs in which the edges probability of occurence is fixed to a given value p [35]. The dependence assumption in this case is: all possible distinct edges are independent of one another. In the previous discussion we stated that the only relevant local structure to the model are those in which all possible ties in it are conditionally dependent. In this case, given that all the possible ties are independent, the possible local structures are related to the single edges Y_{ij}. Hence the Eq.6.3.2 becomes:

$$Pr\left(\mathbf{Y}=\mathbf{y}\right)=\left(\frac{1}{\kappa}\right)exp\left(\sum_{i,j}\eta y_{ij}\right) \tag{6.0}$$

in which, to respect Eq.6.3.2, every set A includes just the single edge Y_{ij}. In other words, every edge Y_{ij} represents a local structure A to which it is associated a parameter η_{ij}. The network statistic $g_A(\mathbf{y})=g_{ij}(\mathbf{y})=y_{ij}$ indicates wether the local structure is observed or not, i.e. if there is the tie or not. To simplify the Eq. 6.3.3 without loss of generality, we can impose the following equality $\eta_{ij}=\theta\ \forall(i,j)\in V$. This equality is called homogeneity and by using it the Eq. 6.3.3 becomes:

$$Pr\left(\mathbf{Y}=\mathbf{y}\right)=\left(\frac{1}{\kappa}\right)exp\left(\theta L(\mathbf{y})\right) \tag{6.0}$$

where $L(\mathbf{y})=\sum_{i,j}y_{ij}$ denotes the number of edges observed in $\mathbf{y}$ and θ is the so-called *edge parameter* related to the probability of of occurrence of a tie.

Dyadic models

When we deal with digraph a little complication can be applied in the model. By means of the dyadic models we introduce that dyads, rather than edges, are indepedent of one another. By this we have two local structure in the model: the single edges and the reciprocated edges. Assuming the homogeneity, as in the Bernoulli models, the specification of the ERGM is:

$$Pr\left(\mathbf{Y}=\mathbf{y}\right)=\left(\frac{1}{\kappa}\right)exp\left(\theta\sum_{i,j}y_{ij}+\rho\sum_{i,j}y_{ij}y_{ji}\right)=\left(\frac{1}{\kappa}\right)exp\left(\theta L(\mathbf{y})+\rho M(\mathbf{y})\right) \tag{6.0}$$

where $L(\mathbf{y})$ is the number of edges in $\mathbf{y}$ and M(y) is the number of mutuyal ties in $\mathbf{y}$. The Eq. 6.3.3 expresses the most simple dyadic independence model. Somewhat more complicated model specification is the p_2 model, in [**?**] and in [58], where it is assumed dyadic independence as well but conditional on node attributes. The p_2 model works when structure is expected to arise from attributes of te actors. Obviously, in the case of undirected networks Bernoulli random graph and dyadic models are identical. Indeed in undirected case, the reciprocity parameter ρ in Eq. 6.3.3 is irrelevant and the model becomes the one expressed by the Eq. 6.3.3.

Markov graphs and p* models

In a 1986 paper Frank and Strauss introduced Markov dependence in order to adapt the ERGM to more realistic situations [36]. In this model, a possible edge between nodes *i* and *j* is assumed to be dependent on any other possible tie that involve *i* or *j*. In this case, the two ties (because we are in a directed graph) between *i* and *j* are said to be

conditionally dependent, given the values of all other ties. In particular, when two ties are conditionally dependent, and a value of one of them changes, the probability of the other tie is influenced, even if all other ties in the network remain the same.

The Markov dependence assumption consists in the fact that two possible edges are conditionally dependent if and only if they are incident with a same actor. If moreover we asume also homogeneity, we obtain the Markov random graph model. This model has several local structures for both the directed and undirected case.

These local structures and associated parameters are related to the discussed structural regularities in a social network, i.e. the local structures (or configurations) censuses [65]. Here are included both the parameter in the Bernoulli random graph (τ_{15}) and in the dyadic model (τ_{11}); moreover there are also several two-star effects: the two-out-star parameter (τ_{12}) that is related to the expansiveness, the two-mixed-star parameter (τ_{13}) relates with the two-paths, and the two-in-star parameter (τ_{14}) realtes to popularity. Also important triadic effects are included in the model: the transitivity (τ_9) and the cyclic triads (τ_{19}). These are not the full set of possible local structures that one can include in the model: for example all the higher order stars. However, except some three-star effect, it is very hard to estimate the model if all such stars are included because it results that there are too may parameters. A Markov graph model for undirected graph with edge, two-star, three-star and triangle effect is:

$$Pr\left(\mathbf{Y}=\mathbf{y}\right)=\left(\frac{1}{\kappa}\right)exp\left(\theta L(\mathbf{y})+\sigma_2 S_2(\mathbf{y})+\sigma_3 S_3(\mathbf{y})+\tau T(\mathbf{y})\right) \tag{6.0}$$

where $S_2(\mathbf{y})$, $S_3(\mathbf{y})$ and $T(\mathbf{y})$ are the numbers of two-star, three-star and triangle, respectively, in the network $\mathbf{y}$. The model in the Eq. 6.3.3 is an example of how it is possible to set to zero some higher order parameters. In this case, we assume that the distribution of stars (namely, the degree distribution) can be completely explained by using only the two-star and three-star effects.

Several SNA applications deal with this important subfamily of ERGM, that is called *triad model*, which more precisely is a specification of Markov graph model in Eq. 6.3.3 [80], [36]. Its simplified probability function, for undirected graphs, is:

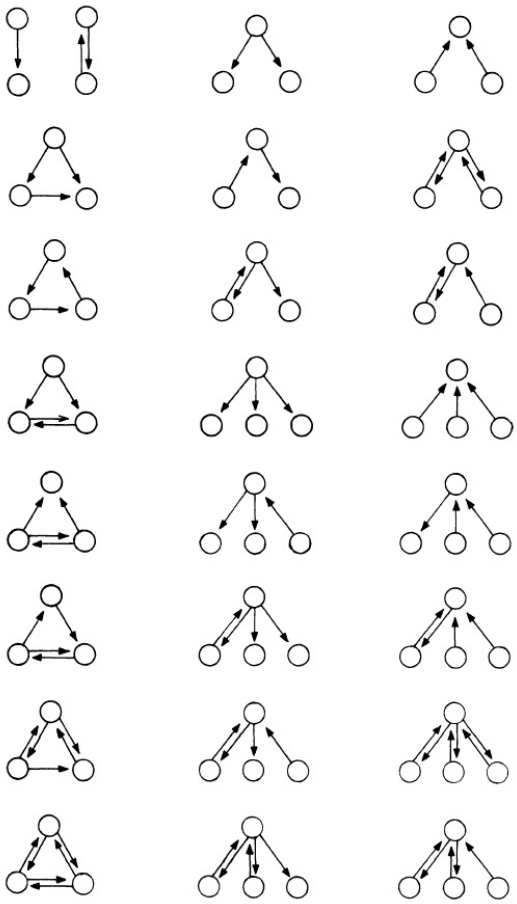


Figure 6.3: The ensemble of sufficient subgraphs (a.k.a. configurations or local structures) of a directed Markov graph of order 4. The counts of these substructures in the observed graph furnish a sufficient statistics for the network parameters (source: Frank, Strauss 1986).

$$Pr_{\vartheta}(\mathbf{Y}=\mathbf{y}) = exp(\vartheta_1 L(\mathbf{y}) + \vartheta_2 S_2(\mathbf{y}) + \vartheta_3 T(\mathbf{y}) - \Psi(\vartheta)) \tag{6.0}$$

where we specified only the parameters $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3)$, which is respectively the "ties", the "two-star" and the "transitivity" parameters; $L(\mathbf{y})$, $S_2(\mathbf{y})$, $T(\mathbf{y})$ are sufficient statistics for the number of ties, for the number of two-stars and for the number of triangles; $\Psi(\vartheta)$ is the normalizing constant under the exponentiation (which is in Eq. the factor $\left(\frac{1}{\kappa}\right)$). The examples in the following are made on this simplified model. However the extension to the general $p*$ model is quite straigthforward.

In particular from the observed adjacency matrix we can compute the values of $L(\mathbf{y})$, $S_2(\mathbf{y})$ and $T(\mathbf{y})$ in this way:

$$\begin{aligned} L(\mathbf{y}) &= \textstyle\sum_{1\le i<j\le n} y_{ij} \\ S_2(\mathbf{y}) &= \textstyle\sum_{1\le i<j\le n} \sum_{k\neq i,j} y_{ik} y_{jk} \\ T(\mathbf{y}) &= \textstyle\sum_{1\le i<j<k\le n} y_{ij} y_{ik} y_{jk} \end{aligned}$$

Obviously for $\vartheta_2 = \vartheta_3 = 0$ we return to the Bernoulli graph model. Of course the triad model can be specified in other simpler submodels depending on which parameter is set to zero: for $\vartheta_2 = \vartheta_3 = 0$ we have the reciprocity model (that refer to the Bernoulli graphs in previous paragraph); for $\vartheta_3 = 0$ we have the two-star model (that refer to the Bernoulli graphs in previous paragraph). The generalisation of Eq. 6.3.3 to arbitrary statistics $u(\mathbf{y})$ (see [87]) leads to the $p*$ family models. Now we can appreciate that $p*$ are ERGM models in Eq. 6.3.2 of the form:

$$Pr_{\vartheta}(\mathbf{Y}=\mathbf{y}) = exp\left(\vartheta' u(\mathbf{y}) - \Psi(\vartheta)\right) \tag{6.0}$$

where $\mathbf{y}$ is the observed adjacency matrix and $u(\mathbf{y})$ represents a generic vector of statistics of a general digraph and ϑ is a vector of network parameters. Some different specifications are made by considering this models focused on the subgraphs counts [39], [62] that here we called local structures.

As we observed in the case of general Markov graphs in Eq.**??**, when we deal with directed networks, the number of possible parameters increases. Indeed, every edge

point in two directions which it means that the number of possible connections between pairs of nodes and triple of nodes must be considered in several ways (see fig. 6.3).

ERGM wiht node attributes

The hypothesis that the social structure depends on individual choices and characteristics it could be included in the ERGM models. Indeed, it is possible to introduce some actor attributes (as gender, education level, etc.) in ERGM and specifically in Markov graphs. There are a number of ways to introduce these variables. The simpler is to define an $\mathbf{X}$ vector of dichotomus attribute variables, in which the *i*-th entry is $X_i = 1$ if the *i*-th actor has the attribute, and $X_i = 0$ otherwise. The vector $\mathbf{x}$ is the vector of observed attributes. The specification of individual variables allows to investigate how exogeneous factors influence the relational ties formation. It is often used for describing the a similarity or homophily hypothesis as a basis for social selection, by looking at the distribution of ties given the distribution of attributes [40]. Formally, the Eq. 6.3.2 becomes a conditional probability distribution of the form $Pr(\mathbf{Y} = \mathbf{y}) \, |\mathbf{X} = \mathbf{x})$.

A simple attribute dependence hypothesis is the one in which the *i*-th node's attribute influences all the possible ties involving *i*, i.e. Y_{ij}. This kind of dependence is usually indicated as *Markov attribute assumption*. However, this dependence seems a realistic assumption for small networks, i.e. networks where the personal attrubute can fully determine all the possible actor's relations.

6.3.4 Estimation of the $p*$ models parameters

The basic estimation procedure for the $p*$ models parameters are the *pseudo-likelihood estimation*, first used for networks by Ikeda and Strauss in [26] in order to estimate the Markov models parameters. Indeed, for Markov random graph models, standard maximum likelihood becomes very untractable in large networks because of the difficulties to compute the normailizing coefficient $\left(\frac{1}{\kappa}\right)$ in Eq. 6.3.2. Because of the strength of dependency constraints, standard statistical procedures cannot directly be applied for ERGM models. However these problems are recently overpassed by the use

of the *Monte Carlo maximum likelihood* techniques.

Pseudo-likelihood estimation

The maximum pseudo-likelihood (PL) approach was first introduced by Besag in [9] for highly interactive data in lattice systems. The pseudo-likelihood estimation is very useful in ERGM because it allows the fit of very complicated models. The waek point of this techinique is however that the properties of the obtained estiamtors are not well-known and, in several cases, the estimates are not accuarate.

PL requires to be applied that the Eq. 6.3.2 must be transformed in the following equivalent *conditional form* (see [26] for a more detailed explaination):

$$log \left[\frac{Pr\left(Y_{ij} = 1 \left| \mathbf{y}_{ij}^{C} \right.\right)}{Pr\left(Y_{ij} = 0 \left| \mathbf{y}_{ij}^{C} \right.\right) = \sum_{A(Y_{ij})} \eta_A \delta_A(\mathbf{y})} \right] \quad (6.0)$$

where: the sum is over all local structures A that contain Y_{ij}; η_A is the parameter corresponding to the local structure A; $\delta_A(\mathbf{y})$ is the *change statistic*, i.e. the change in the value of the network statistic $z_A(\mathbf{y})$ when y_{ij} passes from 0 to 1; $\mathbf{y}_{ij}^{C}$ are all the observations of ties in $\mathbf{y}$ except the observation y_{ij}.

Whenever the change statistic is computed, in the calculation of the pseudo-likelihood estimates, every possible tie Y_{ij} becomes a case in a standard logistic regression procedure, with y_{ij} predicted from a set of change statistics [23].

The procedure illustrate in [23] it is very similar to the logistic regression and to the loglinear models. However, with ERGM we assume dependent observations whereas in logistic regression it is only assumed the independency among them. This is the reason why the parameter estimates may be biased and the standard error may be too small [40]. Another difference between PL procedure and standard logistic regression is that it is not possible to assume that the PL deviance is distributed as a Chi-square, whereas this is true in logistic regression. An obvious limit of npseudolikelihooed methods is that it is not an admissible estimator for a squared-error loss functions [80], [**?**]. Another limitation of PL is that the accuracy of the PL estimates increases when the depedence among observations is not very strong. This is the main weak point of PL.

Indeed, the use of a procedure for dependency structured data that is not accurate when the dependency is too high allowed the development of further estimation approaches.

For these reasons most important of these new approaches is the Markov chain Monte Carlo procedure. We discuss about it in the following section.

Markov chain Monte Carlo maximum likelihood estimation

Markov chain Monte Carlo (MCMC) has become increasingly popular as a general purpose class of approximation methods for complex inference, search and optimization problems. A variety of MCMC samplers (e.g., the Gibbs sampler or the Metropolis samplers)can be constructed for any given problem by varying the sampling distribution subject to conditions that ensure convergence to the desired distribution.

The most important motivation to the development of the general Markov chain Monte Carlo estimation and Markov chain Monte Carlo in maximum likelihood estimation (MCMCMLE) is perhaps that the simulation of these models can be easily implemented [24]. In particular, simulation of the graph distribution for a given set of parameter values can be obtained by using several well-known alghorithms (e.g., the Metropolis algorithm). The simulation techniques in MCMCMLE could be implemented in very different ways but basically there is a common outline to follow: i)simulation of a distribution of random graphs from a set of known parameters by means of a random draw; ii) parameters adjustment by the comparison of the observed network and the generated distribution, until the parameter estimated values have been stabilize (i.e. until the stationariety).

There are several random draws simulator available. The most used is the Metropolis-Hasting sampler, which is a special case of the Metropolis sampler, obtains the state of the chain at the time $t+1$ by sampling a candidate point $\mathbf{Y}*$ from a proposal distribution $q(\cdot|\mathbf{Y}^t)$:

$$\alpha(\mathbf{Y}^t, \mathbf{Y}*) = min\left\{1, \frac{\pi(\mathbf{Y}*)q(\mathbf{Y}^t|\mathbf{Y}*)}{\pi(\mathbf{Y}^t)q(\mathbf{Y}*|\mathbf{Y}^t)}\right\} \tag{6.0}$$

The candidate point $\mathbf{Y}*$ is accepted as the next chain state if its conditional prob-

ability of occurrence is higher than the one obtained by an uniform distribution: $U \leq \alpha(\mathbf{Y}^t, \mathbf{Y}*)$. The major advantage of such a sampler is that the distribution of interest, namely $\pi(\mathbf{Y})$, is obtained as a ratio, so the constant of proportionality, which is the most difficult element to estimate, is canceled out.

Instead, in the case of $p*$ and ERGM, the Gibbs sampler [1] is prefereed [80]. This sampler is applied to the elements of adjacency matrix [80], [73]. This means that an initial matrix $\mathbf{Y}^{(1)}$ is selected, and the elements of this matrix are randomly updated. This update is just a random process that is a Markov chain $\mathbf{Y}^{(t)}$ and its distribution asymptotically tends toward the desired random graph distribution. It is worth to note that two consequent matrices, namely $\mathbf{Y}^{(t)}$ and $\mathbf{Y}^{(t+1)}$, differ at most for just one element, i.e. the element updated in the *t*-th step.

Denoting with Y_{ij} the element updated to the step *t*, which corresponds to the candidate point $\mathbf{Y}*$ in the Metropolis sampler [48], we express its new value as obtained in according to the following conditional distribution (given all the other elements):

$$Pr_\vartheta \left\{ Y_{ij}^{(t+1)} = a \middle| \mathbf{Y}^{(t)} = \mathbf{y} = Pr_\vartheta \left\{ Y_{ij} = a \middle| Y_{hk} = y_{hk} \forall (h,k) \neq (i,j) \right. \right. \tag{6.0}$$

In this step, all other elements are left unchanged: in other terms, $Y_{hk}^{t+1} = Y_{hk}^{t}$ $\forall (h,k) \neq (i,j)$. The term in the right side of Eq. 6.3.4 represent the conditional distribution (that is the same used in PL procedure) and the left is the transition probability that must be defined.

Gibbs sampling is applicable when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. Indeed in SNA, it is used the Gibbs sampler because a theorem in [73] affirms that the distribution of the digraph $\mathbf{Y}^{(t)}$ in the limit of large realizations of the stochastic process, i.e. for $t \to \infty$, converges

[1] The Gibbs sampling is one of several algorithm that in MCMC are used to generate a sequence of samples from a joint probability distribution of two or more random variables. The Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm which is usually faster and easier to use but is less generally applicable. Indeed it is applicable only when the target distribution is known for the single varaiables in the process.

to the exponential random graph distribution in Eq. 6.3.2. Therefore we do not need of a proposal probability distribution $q(\cdot|\mathbf{Y}^t)$ knowing that the target distribution is just the ERGM in Eq. 6.3.2. Hence we just need to adapt the conditional probability in Eq. 6.3.4 in order to obtain the Eq. 6.3.2 (and also the $p*$ model in Eq. 6.3.3).

ERGM, regradless its particular specifications, is generally obtained by a logistic regression where the sufficient statistics (the counting of the configuration in fig. 6.3) are given by difference between the values for $u(\mathbf{y})$ occurred when letting the elements y of $\mathbf{y}$ be $y_{ij} = 0$ and $y_{ij} = 1$, and leaving the other elements equal:

$$logit(Pr_{\vartheta}\left\{Y^{ij=1}\middle|\, Y_{hk} = y_{hk}\right) = \vartheta^{'}(u(\mathbf{y}^{(ij1)}) - u(\mathbf{y}^{(ij0)}))\forall(h,k) \neq (i,j) \tag{6.0}$$

where $\mathbf{y}^{(ij1)}$ and $\mathbf{y}^{(ij0)}$ are the adjacency matrices obtained by defining the elements (i,j) respectively as $y_{ij}^{(ij1)} = 1$ $y_{ij}^{(ij0)} = 0$, and leaving the other elements unchanged [80].

CHAPTER 7

Spectral graph theory in Social Network Analysis

In this chapter we will show how the complex networks models and the spectral graph theory tools could be usefully implemented in real case analysis, especially in dealing with social network data.

Throughout the present dissertation we discussed about network analysis in a very general fashion and we proposed how SNA could be studied in a more general theoretical framework. In this chapter we will show some advantage deriving by applying this scheme. Specifically it is central for us to consider social networks as particular case of complex system in such a way it could be possible the study of their properties by using a more general perspective and a number of different analytic tools.

This approach does not represent a criticism toward the application of the traditional SNA local structures approach usually applied. It is worth to note the importance of several ERGM based results to modelling real networks. In particular the asymptotic properties of the ERGM and the flexibility of p* model are perhaps the main motivations of the applicability of such techniques. Moreover, sometimes the analysis of local configurations of actors it is very convenient because it is also very understandable. In particular the labelling in terms of social behaviour and social positions of such local structures is very appealing.

However we can find some weak points in such a way of studying real social inter-

actions. In particular the propensity to interact in dyads, two-stars, three stars and so on cannot explain the complex propensity of activate/deactivate social relations. Morevoer in ERGM a linear interactions among the local configurations is assumed. Instead for us a social network is a complex system where elementary parts are represented by actors and the single interrelations among them are the basic dynamical components from which emerge the general state of the system. We will apply the methods of spectral graph theory and complex network theory to some SNA problems. In particular, our main research questions are:

- Given k observed configurations of a single social network it is possible to define a comparison between these states in such a way we can measure the difference along discrete time points?

- Given an observed set of relations (our system current state), and defined a cumulative distance in such a way this distance represent actors' relational proximity, it is possible to infer on the possible connections that could be appear/disappear in the next time observation?

These two issues arise because of two main motivations. The first deriving form the fact that the specific problem of network comparison is not widely faced in literature and the few applications. The second is related to the generation of a family of graphs. This method is based on the inference on the possible connections between couple of actors along time, i.e. birth and death of relational ties, given the relational observed current status of the actors. This latter point is also important because the generation of a family of graphs is central in estimation of network parameters.

These two questions are somewhat interrelated because for their solution we implement the spectral graph theory concepts.

7.1 Node distance within networks by using laplacian and random walks on graph

In this section we deals with a very frequent problem in SNA, the comparison between observed networks in a static fashion. This means that we consider just distinct network configurations in different discrete time points. Moreover, we require that the compared configurations must have the same actor set $V = \{v_1, v_2, ..., v_n\}$ of cardinality n. At least we require that exists a one-to-one correspondence between the nodesets of the k graphs.

The main purpose is to compare two network configurations G_i $(i = 1, 2)$, by means of a specification of a real valued relational distance. Such a distance involves the centrality of the actors and it allows to compare the change in node centrality among the different configurations. Here we focus on the comparison between two network states. The extension to k-configurations case is straightforward.

In particular, the method consists of two steps: first, define a distance matrix Δ between nodes of the k networks and second compare these distance matrices. How to build distance matrices on a graph is a not-trivial task. Indeed, in a network the natural induced distance, the geodesic, is not a metric but a quasi-metric (see for instance the section 2.1.4). This fact complicates the projection of the graph in a metric space, as the canonical Euclidean space. We will show that by means of the laplacian, and by using the random walks model on the graph, the passage from a non metric space to a full metric one is possible. This is only one of the important application of the spectral graph quantity for a resolution of a typical network problem.

7.1.1 The network distance problem in SNA literature

In SNA literature, the comparison between network configurations is basically faced using two approaches: i) the combinatorial approach [81]; ii) the local structures censuses approach [34]. We indicate these methods, respectively, with MA and MB. Briefly, the first method (MA) is based on the comparison between the $k = 2$ adjacency matrices related to the k configurations by using classical statistial indices

usually applied for variables (e.g. correlation coeficient, Goodman-Kruskal Gamma, etc.). In particular, in this method, the comparison is made by measuring the observed differences between two adjacency matrices by mean of the chosen index α, that could be for example the correlation coefficient. Consequently, by using a quadratic assignment procedure (QAP) several combinations of the adjacency matrix related to the first observed network are generated. A distribution of the differences $\alpha - \alpha_{comb}$ is made and the observed one is positioned in the corresponding quantile.

The second approach (MB) is made by comparing the frequency distributions of the local structures, in particular triad census, observed in the k configurations [55], [34]. In particular, here is supposed that two network are similar if they share an high number of equal triad isomorphism classes. This method is developed by construct a contingency tables $t \times k$ (where are the 16 triad isomorphism classes) and applying a Correspondence Analysis to that table in order to project the k matrices in a common factorial space and comparing them.

By means of spectral graph theory quantities discussed in chapter 3, in particular the non normalized laplacian matrix and the random walk on graph, we develop a method that is more network oriented than MA and more general to respect MB.

In particular, for the comparison an euclidean node-distance matrix will be build up for each network and the matrices will then be projected in a lower dimensional common space. This allows to detect how the relations observed in E_1 evolve in the ones observed in E_2. This procedure is not dynamically based, but it considers the k configurations as discrete time point observation of a given social network. In particular, we propose to compare network configurations (or states of the network system) defining a continuous relational distance among the nodes for each the *k* states and to detect the differences observed in. The distance we build by this approach is an euclidean distance among the nodes. This allows to compare configurations by means of the projection in a common euclidean space or by using statistical approach for quantitative data.

The approach consists in two steps: i) definition of the distance matrix Δ_i with $i = 1, ..., k$ on the nodes involved in the configurations; ii) comparison between these

matrices.

7.1.2 Definition of the distance matrix on a network

Let V be a set of *n* actors $\{v_1, v_2, \ldots, v_n\}$ and E_1 the set of m_1 observed unordered couples of elements of V indicate with $\{e_{11}, e_{12}, \ldots, e_{1k}\}$, then $G_1(V, E_1)$ is the network generated by V and E_1. Let E_2 be a different set of m_2 observed edges $\{e_{21}, e_{22}, \ldots, e_{2h}\}$ generating the network $G2(V, E_2)$. Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be the adjacency matrices associated respectively to the networks G_1 and G_2 and $\mathbf{D}_1$ and $\mathbf{D}_2$ be the $n \times n$ diagonal degree matrices, i.e. the matrix in which the diagonal elements d_{ii} are the degrees of *i*-th node. Therefore, in the following discussion, we consider only networks represented by simple unweighted graphs. However, the generalization to the weighted case is immediate.

The laplacian and its pseudo-inverse (see sections 3.2.1 and 3.2.2) can be used in finding useful node-distances on a graph. However, in order to define the node-distances we need to consider that, on our networks, is defined a Random Walk (see section 5.3). In general, this kind of model is very useful for social networks, because there is not a loss of generality to consider that for an actor the activation of relationships mainly depends on its neighbours. In particular we will show how the distance we will define is based on the frequency and on the length of the paths among the nodes: this means that two actors are closer bacause of their proximity (if they are connected) but also if they share several closer friends. Briefly, we aim at defining a relational based distance.

In particular, in order to define this distance, in the following we will introduce several types of transition related quantities based on random walk. The central point here is that these quantitites are strictly related to the laplacian pseudo invesre $\mathbf{L}^+$.

The inverse laplacian is a matrix that incapsulate the property of the transition matrix $\mathbf{P}$ corresponding to the random walk on the graph (see chapter 5). Moreover, $\mathbf{L}^+$ to respect with the matrix $\mathbf{P}$ has the useful charcateristic of being doubly stochastic. This means that its entries could represent probabilties of a particular form. Moreover, another property is that its quadratic transformation furnish a distance between nodes. This distance represents the time that a random walker needs to pass through the various

states taking into account the commutation of the starting point (coming back to its starting point).

The first quantity, defined on RW, is the *avarage first passage time* $m(j\,|i)$ wich represent the mean number of steps a random walker needs to reach, for the first time, the node *j* starting form the node *i* [53]. More precisely this quantity measure the minimum time until hitting state *j* as $T_{ij} = min(t \geq 0\,|s(t) = j, s(0) = i)$.

It is possible to prove that this quantity is directly obtained by the (i, j)-th element l_{ij}^{+} of $\mathbf{L}^{+}$ [**?**]:

$$m(j\,\Big|\,i) = \sum_{k=1}^{n}(l_{ik}^{+} - l_{ij}^{+} - l_{jk}^{+} + l_{jj}^{+}) \quad (7.0)$$

where: *k*= node(s) between *i* and *j*.

The second related quantity is the *average commute time* (CT), that is defined from $m(j\,|i)$. CT represents the mean number of steps a random walker have to use to reach, for the first time, the node *j* from *i* and coming back to the node *i* [60]:

$$C\bar{T}(ij) = m(j\,|i) + m(i\,|j) \quad (7.0)$$

It is easy to show that also this quantity it is easily obtained by the $\mathbf{L}^{+}$ elements (cfr. sect3.2.2). Its average form is:

$$C\bar{T}(i, j) = V_G(l_{ii}^{+} + l_{jj}^{+} - 2l_{ij}^{+}) \quad (7.0)$$

where: V_G is the volume of the graph (i.e. the sum of all the degrees).

It can be showed that this quadratic form Eq. 7.1.2 is a distance because it satisfies all the metric axioms [33] [3].

Then we can rewrite Eq. 7.1.2 by considering that the *i*-th unit basis vector represents the *i*-th node in the graph:

$$ct(i, j) = V_G(\mathbf{e}_i - \mathbf{e}_j)^t\mathbf{L}^{+}(\mathbf{e}_i - \mathbf{e}_j) \quad (7.0)$$

In particular the distance between two single nodes is expressed by the following

formula:

$$CT(i,j) = (l_{ii}^{+} + l_{jj}^{+} - 2l_{ij}^{+}) \tag{7.0}$$

The principal property of the average commute time distance is that its square root is a distance in the canonical n-dimensional Euclidean space $\Re^n$ of the nodes, called *Euclidean commute-time Distance* or ECTD [**?**]:

$$ectd(i,j) = \sqrt{ct(i,j)} \tag{7.0}$$

It is easy to show that the ECTD in Eq. 7.1.2 is a Mahalanobis distance with weight matrix $\mathbf{L}^{+}$.

The useful characteristic of ECTD is that it decreases when the number of walks increases and/or the length of these walks decreases. In short, if between two actors there is a low ECTD they can be considered close to each other, given that they have *many short paths connecting them*. In SNA terms this have nicely consequences: it could be interpret as the probability that an actor i has a relation with j considering both the relational status, in terms of degree centrality and closeness, of i and j.

It is straigthforward that for k compared matrices we could obtain k node-distance matrices $\Delta_i \; i = 1, ..., n$ of distance among the nodes of the graph, in which the elements are the distances ECTD between every pairs of nodes (*i*,*j*) in $\Re^n$. Moreover, the property of this distance of being euclidean, allows us to implement some classical dimensionality techniques to reduce dimensionality, e.g. Multidimensional Scaling [50].

Considering our comparison problem. We can now obtain $k = 2$ ECTD distance matrices, Δ_1 and Δ_2 respectively for the network G_1 and the network G_2.

7.1.3 Comparing distance matrices in a common euclidean space

We showed how using some spectral graph theory quantity and defining a particular model on our network it is easy to discover very interesting property. Now the next step, therefore, consists in the reduction of the dimensionality of the distance matrices

in order to graphically represent the networks G_1 and G_1 in $\Re^n$ by means of some special dimensionality reduction statistical techniques. In particular we will use *metric Multidimensional Scaling* (see for instance: Borg and Groenen).

In order to compare the networks it is crucial to note that, as we already stated, ECTD is a distance between two nodes that decrease when increase the number of connections between these nodes or when the length of such connections decrease. The interpretation criteria of the two network's differences is an elementwise comparison between the nodes in G_1 and the corresponding nodes in the configuration G_2. In particular, considering the individual positions of the nodes in the two compared networks. In fact, we interpret ECTD distance as a relative distance of the *i*-th node to respect all the other nodes in the network. Therefore a variation of this distance implies a modification of the position of *i* to respect the whole network.

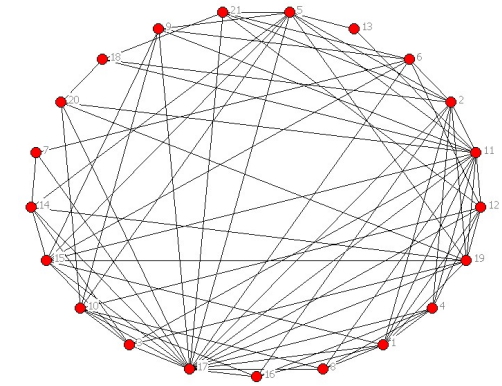
7.1.4 Example network application

In this section we illustrate how the method works. Here we consider a simulated undirected network G_1 and G_2 on $N = 21$ nodes (see fig.7.1) with adjacency matrices reported in the plot.

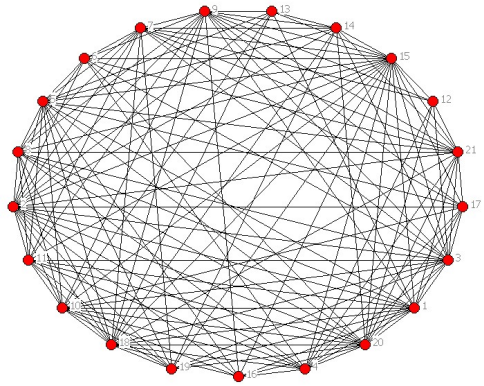
We simulated networks in such a way G_2 is derived, by a stochastic process, from the network G_1. In an hypothetical time flows, G_2 follows G_1. In particular, the networks are obtained as random graph $\mathcal{G}_{21,p}$ with different probabilities p. In particular we set $p = 0.6$ for G_1 and $p = 0.75$ for G_2, in such a way we obtain a more connected network in the second configurations. By this we will show how the ectd distance is higher in $G2$ rather than in G_1. We will explain the method execution and the computation of the distance matrices.

As we anticipated, by using different probabilities in the family opf random graph models we differentiate the probability of obtaining a link between couple of nodes. Indeed it is clear by observing the degree distributions of G_1 and G_2 in the table that network G_2 has an higher connectivity.

By using Eq.3.5 we obtain the combinatorial laplacian associated to the two network configurations. In table 7.2 it is showed the laplacian of G_1. It is possible to observe

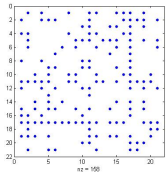


(a) G_1

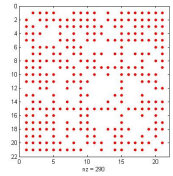


(b) G_2

Figure 7.1: (a) The initial network configuration G_1 (b) The network configuration G_2 at the state 2.



(a) $\mathbf{A}_1$



(b) $\mathbf{A}_2$

Figure 7.2: (a) Plot of the entries of the adjacency matrix $\mathbf{A}_1$ of the initial network configuration G_1 (b) Plot of the entries of the adjacency matrix $\mathbf{A}_1$ network configuration G_2.

Table 7.1: Degree distribution of the network in the two configuration state G_1 and G_2 (see fig.7.2).

Degree	G_1	G_2
v_1	9	16
v_2	10	19
v_3	6	17
v_4	7	14
v_5	10	16
v_6	7	10
v_7	3	14
v_8	5	14
v_9	6	16
v_{10}	8	15
v_{11}	14	13
v_{12}	8	7
v_{13}	2	8
v_{14}	6	11
v_{15}	9	20
v_{16}	5	9
v_{17}	18	11
v_{18}	4	17
v_{19}	10	11
v_{20}	5	17
v_{21}	6	15

that the row sum is equal to zero.

The pseudo-inversion in the sense of the Eq. 3.10 of the laplacian will furnish the pseudo-inverse $\mathbf{L}^+$. In table 7.3 we report the pseudo-inverse $\mathbf{L}_1^+$ for the laplacian $\mathbf{L}_1$ of the graph G_1. It is possible to observe that the row sum and the column sum are now is equal to one. This means that $\mathbf{L}^+$ is a matrix that contains the transition probabilities of the matrix $\mathbf{P} = \mathbf{DL}$ but in a more "relational sense".

Indeed by using its elements now we can compute the related commute-time distances for the two networks. The computation is made by Eq.7.1.2. In the table 7.4 we will show the node distances among the nodes $\mathbf{CT}_1$ in network configuration G_1. The values of $ct(i, j)$ express the fact that between these two nodes there are a number of short path connecting them.

Table 7.2: Laplacian matrix $\mathbf{L}_1$ of the network G_1

Nodes	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{16}	v_{17}	v_{18}
v_1	9	-1	0	-1	0	0	0	-1	0	0	-1	-1	0	0	-1	-1	-1	0
v_2	-1	10	0	-1	-1	-1	0	0	0	0	-1	0	0	0	0	-1	-1	-1
v_3	0	0	6	0	0	0	0	0	0	-1	-1	0	0	-1	-1	-1	0	-1
v_4	-1	-1	0	7	0	0	0	-1	0	0	-1	-1	0	0	-1	0	-1	0
v_5	0	-1	0	0	10	0	0	0	-1	-1	-1	0	-1	-1	-1	0	-1	0
v_6	0	-1	0	0	0	7	-1	0	-1	0	0	-1	0	0	-1	0	-1	0
v_7	0	0	0	0	0	-1	3	0	0	0	0	0	0	-1	0	0	-1	0
v_8	-1	0	0	-1	0	0	0	5	0	-1	-1	0	0	0	0	0	-1	0
v_9	0	0	0	0	-1	-1	0	0	6	-1	-1	0	0	0	-1	0	-1	0
v_{10}	0	0	-1	0	-1	0	0	-1	-1	8	0	-1	0	0	0	-1	-1	0
v_{11}	-1	-1	-1	-1	-1	0	0	-1	-1	0	14	-1	-1	0	-1	0	-1	-1
v_{12}	-1	0	0	-1	0	-1	0	0	0	-1	-1	8	0	0	0	0	-1	0
v_{13}	0	0	0	0	-1	0	0	0	0	0	-1	0	2	0	0	0	0	0
v_{14}	0	0	-1	0	-1	0	-1	0	0	0	0	0	0	6	-1	0	-1	0
v_{15}	-1	0	-1	0	-1	-1	0	0	-1	0	0	0	-1	9	0	0	-1	0
v_{16}	-1	-1	0	-1	0	0	0	0	0	-1	0	0	0	0	0	5	-1	0
v_{17}	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	18	0
v_{18}	0	-1	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	4
v_{19}	-1	-1	-1	0	-1	0	0	0	0	0	-1	-1	0	-1	-1	0	-1	0
v_{20}	0	0	0	0	0	0	0	0	0	-1	-1	0	0	0	0	0	-1	-1
v_{21}	0	-1	0	0	-1	-1	0	0	0	0	0	-1	0	0	0	0	-1	-1

If we focus on table 7.4 on the node v_{17} we see that is the node with shorter distance in average. This is due to the fact that it is the more connected one, its degree is equal to 18, which is the larger value. It is also central in the closeness sense because is very reachable from everyone in the network. Moreover node v_7 is the further from the others. However, its distance is not very high in average because it is connected to the node v_{17} that share an high number of ties. Thus, the actor v_7 can be connected with the other nodes because of its undirected relations. Finally, let consider the larger value in this part of the table, which is $ct(7, 13) = 0.6240$. Indeed, if we look at the fig. **??** we can see that the "opportunity" of connection between v_7 and v_{13} are very low because they do not share several common connections. The impotant element is that the commute-time distance indicates the centrality in networks of a node and,

Table 7.3: Part of the pseudo-inverse $\mathbf{L}_1^+$ of the laplacian matrix $\mathbf{L}_1$ of the network G_1

Nodes	v_1	v_2	v_3	v_4	v_5	v_6	$v_7 \ldots$	
v_1	0.1462	0.0500	0.0381	0.0569	0.0381	0.0374	0.0300	...
v_2	0.0500	0.1360	0.0358	0.0515	0.0462	0.0483	0.0323	...
v_3	0.0381	0.0358	0.1876	0.0345	0.0410	0.0354	0.0356	...
v_4	0.0569	0.0515	0.0345	0.1722	0.0357	0.0359	0.0286	...
v_5	0.0381	0.0462	0.0410	0.0357	0.1358	0.0393	0.0337	...
v_6	0.0374	0.0483	0.0354	0.0359	0.0393	0.1714	0.0644	...
v_7	0.0300	0.0323	0.0356	0.0286	0.0337	0.0644	0.2953	...
v_8	0.0563	0.0385	0.0363	0.0607	0.0354	0.0323	0.0277	...
v_9	0.0368	0.0376	0.0395	0.0347	0.0519	0.0545	0.0348	...
v_{10}	0.0400	0.0376	0.0513	0.0401	0.0467	0.0362	0.0298	...
v_{11}	0.0477	0.0462	0.0469	0.0479	0.0466	0.0374	0.0300	...
v_{12}	0.0512	0.0422	0.0377	0.0528	0.0387	0.0510	0.0332	...
v_{13}	0.0286	0.0308	0.0293	0.0278	0.0608	0.0256	0.0212	...
v_{14}	0.0356	0.0351	0.0599	0.0316	0.0505	0.0391	0.0692	...
v_{15}	0.0476	0.0392	0.0549	0.0363	0.0496	0.0499	0.0380	...
v_{16}	0.0567	0.0535	0.0345	0.0613	0.0353	0.0341	0.0281	...
v_{17}	0.0469	0.0457	0.0470	0.0469	0.0452	0.0470	0.0477	...
v_{18}	0.0342	0.0548	0.0307	0.0342	0.0357	0.0348	0.0247	...
v_{19}	0.0489	0.0468	0.0534	0.0390	0.0481	0.0379	0.0343	...
v_{20}	0.0363	0.0385	0.0382	0.0347	0.0370	0.0322	0.0278	...
v_{21}	0.0368	0.0533	0.0325	0.0367	0.0487	0.0560	0.0337	...

considered pairwise can express the relational distance among paiurs of actors.

In table 7.5 we will show the node distances among the nodes $\mathbf{CT}_2$ in network configuration G_2. Here the fact that connectivity is increased led to a great reduction of the commute-time distance among nodes. In average the nodes in G_2 are closer in a CT sense. Moreover the connectivity is very high and therefore the CT converges toward a limiting value for each node in the network.

By the comparison of the distance matrices between the two configurations we can obtain information on the relational role of the nodes. This allows the comparison elementwise. However we can use some average value to resume the differences in distance among the compared networks.

Table 7.4: Part of the commute time distance between the nodes of the network G_1

Nodes	v_1	v_2	v_3	v_4	v_5	v_6	$v_7 \ldots$	
v_1	0	0.1823	0.2576	0.2047	0.2059	0.2428	0.3816	...
v_2	0.1823	0	0.2520	0.2053	0.1794	0.2108	0.3668	...
v_3	0.2576	0.2520	0	0.2908	0.2416	0.2883	0.4118	...
v_4	0.2047	0.2053	0.2908	0	0.2367	0.2717	0.4104	...
v_5	0.2059	0.1794	0.2416	0.2367	0	0.2287	0.3637	...
v_6	0.2428	0.2108	0.2883	0.2717	0.2287	0	0.3380	...
v_7	0.3816	0.3668	0.4118	0.4104	0.3637	0.3380	0	...
v_8	0.2448	0.2702	0.3261	0.2620	0.2762	0.3178	0.4511	...
v_9	0.2592	0.2474	0.2952	0.2894	0.2187	0.2491	0.4123	...
v_{10}	0.2224	0.2171	0.2412	0.2482	0.1987	0.2552	0.3919	...
v_{11}	0.1619	0.1547	0.2049	0.1876	0.1537	0.2076	0.3465	...
v_{12}	0.1991	0.2070	0.2675	0.2218	0.2137	0.2246	0.3843	...
v_{13}	0.4601	0.4455	0.5002	0.4877	0.3853	0.4914	0.6240	...
v_{14}	0.2656	0.2564	0.2585	0.2998	0.2255	0.2838	0.3475	...
v_{15}	0.1966	0.2032	0.2234	0.2452	0.1821	0.2171	0.3649	...
v_{16}	0.2449	0.2411	0.3307	0.2617	0.2772	0.3151	0.4512	...
v_{17}	0.1492	0.1413	0.1903	0.1751	0.1422	0.1741	0.2966	...
v_{18}	0.3247	0.2732	0.3731	0.3508	0.3113	0.3487	0.4927	...
v_{19}	0.1843	0.1782	0.2167	0.2301	0.1755	0.2314	0.3626	...
v_{20}	0.2845	0.2698	0.3220	0.3137	0.2726	0.3178	0.4507	...
v_{21}	0.2611	0.2179	0.3111	0.2873	0.2270	0.2480	0.4164	...

Finally by making the square root of the elements in $\mathbf{CT}_1$ we obtain te correspondent Euclidean commute time distance $\mathbf{\Delta}_{i=1,2}$, which is our expected distance function. ECTD is a scales the values of $Ct(i,j)$ in such a way that it is possible to project the nodes in an *M*-dimensional euclidean space. Given the restriction that we assumed, the nodeset is equal for the two configurations, we can project the node-points in a *common N*-dimensional euclidean space. In this space, the N nodes are exactly separeted by the ECTD distance among them. Thus it is possible to interpret the node distances as in the usual space $\Re^N$, by using the euclidean 2-norms of the vectors.

Techniques as multidimensional scaling allows us to reduce the dimensionality of the distance matrices and to projecting k matrices in a common space. This latter point

Table 7.5: Part of the commute time distance between the nodes of the network G_2

Nodes	v_1	v_2	v_3	v_4	v_5	v_6	v_7	
v_1	0	0.1039	0.1102	0.1200	0.1124	0.1528	0.1283	...
v_2	0.1039	0	0.1013	0.1122	0.1037	0.1349	0.1122	...
v_3	0.1102	0.1013	0	0.1166	0.1159	0.1395	0.1163	...
v_4	0.1200	0.1122	0.1166	0	0.1277	0.1497	0.1359	...
v_5	0.1124	0.1037	0.1159	0.1277	0	0.1436	0.1205	...
v_6	0.1528	0.1349	0.1395	0.1497	0.1436	0	0.1490	...
v_7	0.1283	0.1122	0.1163	0.1359	0.1205	0.1490	0	...
v_8	0.1199	0.1115	0.1165	0.1268	0.1202	0.1476	0.1273	...
v_9	0.1130	0.1043	0.1094	0.1268	0.1188	0.1433	0.1198	...
v_{10}	0.1146	0.1075	0.1139	0.1237	0.1159	0.1571	0.1325	...
v_{11}	0.1238	0.1161	0.1211	0.1321	0.1246	0.1645	0.1320	...
v_{12}	0.1892	0.1796	0.1712	0.1811	0.1905	0.2179	0.1810	...
v_{13}	0.1616	0.1541	0.1713	0.1840	0.1613	0.2085	0.1828	...
v_{14}	0.1446	0.1276	0.1334	0.1554	0.1355	0.1773	0.1424	...
v_{15}	0.1020	0.0933	0.0988	0.1096	0.1019	0.1329	0.1096	...
v_{16}	0.1500	0.1433	0.1591	0.1577	0.1514	0.1950	0.1723	...
v_{17}	0.1351	0.1274	0.1323	0.1424	0.1361	0.1756	0.1432	...
v_{18}	0.1089	0.1007	0.1069	0.1179	0.1091	0.1490	0.1179	...
v_{19}	0.1358	0.1275	0.1327	0.1539	0.1351	0.1786	0.1429	...
v_{20}	0.1095	0.1012	0.1061	0.1160	0.1095	0.1401	0.1236	...
v_{21}	0.1173	0.1082	0.1123	0.1233	0.1170	0.1444	0.1226	...

is made by using the procrustean rotation that allows to optimally adapt the projection of the point-clouds related to the compared networks. Also principal componenet analysis could be implemented given that now we deal with real valued data [42].

7.2 Network discrete time point evolution

In this section we show an application of the ECTD node distance by which we can check the possibility of generating a family of networks by the activation/deactivation of links between couples of actors during a network discrete time-evolution. Therefore we will deal again with networks with the same number of nodes N. In particular, here

we do not investigate network growth but we try to predict if starting from a network state G_1, a subsequent state G_k is likely to happen or not.

We start from the consideration that two unconnected actors involved in relationships with quite the same individuals are likely to activate a relationship. Similarly, for two connected actors that share few links with the same individuals it is likely to happen the opposite, i.e. it is possible that the link between them could ceases to exist. Therefore, we assume that the actor's distance is low, not just in the sense that they are connected by a short path (in term of geodesic), but also in the sense that they have different opportunities to estabilish a connection. As we showed in the previous section this kind of relational distance between actor pairwise, is represented by means of the ECTD distance among them.

7.2.1 Procedure for generating family of depending graphs

We start from an observed network on N nodes at time t_0 denoted with $G^{(0)}(V, E^{(0)})$ or by its adjacency matrix $\mathbf{A}^{(0)}$ and another state of the same network detected in a different time $G^{(k)}(V, E^{(k)})$.

In order to specify the possible following configurations of this network along time, we adopt the point of view of the actors, specifying the probability of connection among them in terms of their relational distance. From section 7.1 we can useful adopt the ECTD distance among nodes. In particular, we focus on the generation of new links but also on the deactivation of old links. More precisely, starting from the initial configuration G_0, let i and j two disconnected actors at the time $t^{(0)}$ (formally $(i, j) \notin E^{(0)}$) we will affirm that they are likely to be connected at the time $t^{(1)}$ if they have a low ECTD distance at the previous time $t^{(0)}$, given all the other connections. Similarly, let i' and j' two connected actors at the time $t^{(0)}$ (formally $(i, j) \in E^{(0)}$), they are likely to be disconnected at the time $t^{(1)}$ if they have an high ECTD distance at the time $t^{(0)}$, given all the other connections.

Similarly to MCMC, in this method the configuration at the different time points can be considered as the state of a Markov chain, because the network form only depends on the current state. The transition to the state $G^{(1)}(V, E^{(1)})$ therefore it is characterized

by the presence of a new link or the absence of an old link. As well as the MCMC procedure also here we consider the transition of the system states in terms of single elements updates . Thus, the adjacency matrices at two consequential time points $\mathbf{A}^{(t)}$ and $\mathbf{A}^{(t+1)}$ differ from each other by just one entry. Differently from MCMC we do not consider random draw from a proposal distribution but we are guided by the observed ECTD distribution in the selection of candidate nodes i and j. In particular we have two class of couples of candidate nodes: two unconnected nodes and two connected nodes at the time t. Given that they are unconnected $a_{ij}^{(t)} = 1$, the probability of being disconnected at the time $t + 1$ is tested by using the loglikelihood ratio from a logistic regression where the independent variable is the observed ECTD distance and the and we use the observation of the connection realized in a given state k.

Briefly, starting at $t^{(0)}$ from our observed network $G^{(0)}$ we will move to next state by using this procedure:

- compute the $ECTD^{(0)}$ among the nodes of $G^{(0)}$ distance as showed in the section 7.1.4;
- selecting all the unconnected node in $G^{(0)}$ and measuring their ECTD distance;
- the unconnected couple $(i, j)^{(+)}$ with the smaller ECTD will be the candidate link $e_{ij}^{(+)}$ that could be appear in the following state, that we then indicate with $G^{(1)} = G_{(+)}^{(1)}$;
- selecting also all the connected node in $G^{(0)}$ and measuring their ECTD distance;
- the connected couple $(i, j)^{(-)}$ with the larger ECTD will be the candidate link $e_{i'j'}^{(-)}$ that could be disappear in the following step of the chain $G^{(1)} = G_{(-)}^{(1)}$;
- compute the likelihood ratio $a = \frac{\rho(G_{(+)}^{(1)})}{\rho(G_{(-)}^{(1)})}$
  if $a \geq 1$ then we accept transition $G^{(1)} = G_{(+)}^{(1)}$;
  if $a < 1$ then we accept transition $G^{(1)} = G_{(+)}^{(1)}$ with probability a and rejct it with probability $1 - a$, i.e. we accept transition $G^{(1)} = G_{(-)}^{(1)}$ with probability $1 - a$;
- iterate this procedure until the time k or until convergence.

In particular ρ is a function of the ECTD distance. Conveniently we can define $\rho(G)$ as a logistic regression with indepedent variable the ECTD distance at current time [1]. In practice, given the observed state $G^{(k)}$ and the related observed connections, we will use the likelihood ratio between the two choices (adding or removing a link), to decide which is the most probable configuration in the following state of the evolution process. In other words, the process of removing/adding a link it is decided based on the connectivity characteristics of the actors and on the observed edges in t^k. In particular the β parameters of the logistic regression represent the influence of the ECTD in the evolution of the links through the time.

Finally, by defining a procedure to generate a family of networks we can implement a way of generate also distributions of network indices in a different way with respect the classical MCMC approach.

[1] For example in the case of comparison issues we can use the final state of the network as observed values of connection for logistic regression.

Bibliography

[1]

[2]

[3]

[4] R. Albert A.L. Barabasi. Emergence of Scaling in Random Networks. *Science*, 286:509–516, 1999.

[5] J. Hanson B. Hillier. *The Social Logic of Space*. Cambridge University Press, cambridge, UK, 1984.

[6] A.L. Barabasi. *Linked: the New Science of Networks*. Perseus Publishing, Cambridge, MA, 2002.

[7] M.A. Beauchamp. An improved index of centrality. *Behav. Sci.*, 10:161–163, 1965.

[8] J. Besag. Spatial interaction and the statistical analysis of lattice system. *Journal of Royal Statistical Society, series B*, 36:96–127, 1974.

[9] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

[10] G. Bianconi and A. Capocci. Number of loops of size h in growing scale-free networks. *Phys. Rev. Lett.*, 90(7):078701, Feb 2003.

[11] J. Bibby. Generalized inverse matrices. *Journal of Royal Statistical Society, series A (General)*, 135(4):608–609, 1972.

[12] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

[13] Marián Boguñá and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E*, 66(4):047104, Oct 2002.

[14] B. Bollobás. Degree sequences of random graphs. *Discrete Mathematics*, 33(1):1–19, 1981.

[15] B. Bollobás. *Modern Graph Theory*. Springer-Verlag, New York, 2001.

[16] B. Bollobás. *Random Graphs*. Cambridge University Press, Cambridge, UK, second edition, 2001.

[17] P.F. Bonacich. Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92:1170–1182, 1987.

[18] R. L. Breiger. *Explorations in Structural Analysis: Dual and Multiple Networks of Social Structure*. Garland Press, New York, 1991.

[19] R.S. Burt. Positions in networks. *Soc. Forces*, 55:93–122.

[20] F. Capra. *The Hidden Connections*. Flamingo, London, 2003.

[21] Boyer C.B. *A History of Mathematic*. John Wiley and Sons, New York, 1968.

[22] F. Chung. *Spectral Graph Theory*. AMS, New York, 1997.

[23] B. Crouch C.J. Anderson, S. Wasserman. A $p*$ primer: Logit models for social networks. *Social Networks*, 21:37–66, 1999.

[24] E.A. Thompson C.J. Geyer. Constrained monte carlo maximum likelihood for dependent data. *Journal of Royal Statistical Society, series B*, 54:657–699, 1992.

[25] N. Crossley. The new social physics and the science of small world networks. *The Sociological Review*, 53(2):351–358, 2005.

[26] M. Ikeda D. Strauss. Pseudo-likelihood Estimation for Social Networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

[27] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[28] R. Diestel. *Graph Theory*. Springer-Verlag, New York, second electronic edition edition, 2000.

[29] E. Durkheim. *The Division of Labor in Society*. The Free Press, New York, 1933.

[30] E.R. Canfield E.A. Bender. The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theory, Ser. A*, 24(3):296–307, 1978.

[31] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[32] P. Erdős and A. Rényi. On the Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[33] A.A. Jagers F. Gobel. Random walks on graphs. *Stochastic Processes and Their Applications*, 2:311–336, 1974.

[34] K. Faust. Comparing social networks: Size, density and local structures. *Metodoloski Zvezki*, 3(2):185–216, 2006.

[35] O. Frank and K. Nowicki. Exploratory statistical analysis of networks. *Annals of Discrete Mathematics*, 55:349–366, 1993.

[36] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.

[37] L.C. Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, 40:35–41, 1977.

[38] E. Friedgut. Sharp thresholds for graph properties, and the *k*-sat problem, with an appendix by Jean Bourgain. *J. Amer. Math. Soc.*, 12:1017–1054, 1999.

[39] S. Wasserman G. Robins, P. Pattison. Logit models and logistic regression for social network, iii. valued relations. *Psychometrika*, 64:371–394, 1999.

[40] Y. Kalish D. Lusher G. Robins, P. Pattison. An introduction to exponential random graph (p*) models for social network analysis. *Social Networks*, 29:173–191, 2007.

[41] N.L. Johnson G. Simons. On the Convergence of Binomial to Poisson Distributions. *Annals of Mathematical Statistics*, 42:1735–1736, 1971.

[42] M. Gherghi and N.C. Lauro. *Appunti di analisi dei dati multidimensionali*. Napoli, rocco curto editore edition, 2004.

[43] E.N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

[44] C. Godsil and G. Royle. *Algebraic Graph Theory*. New York, second electronic edition edition, 2001.

[45] Z.N. Oltvai A.L. Barabási H. Jeong, B. Tombor. The Large-scale Organization of Metabolic Networks. *Nature*, 407:651–654, 2000.

[46] F. Harary. *Graph Theory*. Addison-Wesley, Reading, MA, 1969.

[47] F. Harary and J.A. Barnes. Graph Theory in Network Analysis. *Social Networks*, 2(5):235–244, June 1983.

[48] W.K. Hastings. Monte Carlo Sampling Methods Using Markov Chain and their Applications. *Biometrika*, 57:97–109, 1970.

[49] P. Holme. *Form and Function of Complex Networks*. PhD thesis, Department of Physics, Umeå University, Umeå, Sweden, 2004.

[50] P.J.F. Groenen I. Borg. *Modern Multidimensional Scaling: Theory and Applications*. New York.

[51] M. Lässig J. Berg. Correlated random networks. *Phys. Rev. Lett.*, 89:228701, 2002.

[52] U.S.R. Murty J.A. Bondy. *Graph Theory with Applications*. New York.

[53] J.L. Snell J.G. Kemeny. *Finite Markov Chains*. Springer-Verlag, 1976.

[54] S. Jukna. *Extremal Combinatorics with Application in Computer Science*. Springer-Verlag, New York, 2001.

[55] J. Skvoretz K. Faust. Comparing networks across space and time, size and species. *Sociological Methodology*, 32:267–299, 2002.

[56] J.H. Powell L. Katz. Measurement of the Tendency Toward Reciprocation of Choice. *Sociometry*, 19:403–409, 1955.

[57] E.K. Lloyd, N.L. Biggs, and R.T. Wilson. *Graph Theory*. Clarendon Press, Oxford, U.K., 1976.

[58] B.J.H. Zijlstra M.A.J. Van Duijn, T.A.B. Snijders. p_2: a Random Effects Model with Covariates for Directed Graphs. *Statistica Neerlandica*, 58:234–254, 2004.

[59] U. Zwick N. Alon, R. Yuster. Finding and counting given length cycles. *Algorithmica*, 17:354–364, 1997.

[60] J. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

[61] S.A. Boorman P. Arabie and P.R. Levitt. Constructing blockmodels: How and why. *Mathematical Journal of Psychology*, 17:21–63, 1978.

[62] S. Wasserman P.E. Pattison. Logit models and logistic regressions for social networks. ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52:169–194, 2007.

[63] R. Muhamad P.S. Dodds and D.J. Watts. An Experimental Study of Search in Global Social Networks. *Science*, 301:827–829, 2003.

[64] S. Leinhardt P.W. Holland. A Method for Detecting Structure in Sociometric Data. *American Journal of Sociology*, 70:492–513, 1970.

[65] S. Leinhardt P.W. Holland. Local structure in social networks. In D.R. Heise, editor, *Sociological Methodology 1976*, pages 1–45. Jossey-Bass, San Francisco, 1975.

[66] S. Leinhardt P.W. Holland. An Exponential Family of Probability Distribution fors for Directed Graphs (with discussion). *Journal of the American Statistical Association*, 76:33–65, 1981.

[67] J. Marco-Blanco M. Romance R. Criado, B. Hernández-Bermejo. Asymptotic estimates for efficiency, vulnerability and cost for random networks. *Journal of Computational and Applied Mathematics*, 204(1):166–171, 2007.

[68] A. Vespignani R. Pastor-Satorras, A. Vazquez. Dynamical and correlation properties of the Internet. *Physical Review Letters*, 87:258701, 2001.

[69] A. Rapoport R. Solomonoff. Connectivity of Random Nets. *Bull. Math. Biophysics*, 13:107–117, 1951.

[70] A. Rapoport. Spread of information through a population with socio-structural bias: Iii. suggested experimental procedures. *Bulletin of Mathematical Biology*, 1:75–81, 1954.

[71] F.S. Roberts. *Discrete Mathematical Models*. Prentice-Hall, Englewood Cliffs, NJ, 1976.

[72] L. Page S. Brin. The anatomy of a large-scale hypertextual websearch engine. *Comput. Netw.*, 30:107–117.

[73] D. Geman S. Geman. Stochastic relaxation, gibbs distributions, and bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[74] A. N. Samukhin S. N. Dorogovtsev. Mesoscopics and fluctuations in networks. *Physical Review E*, 67:037103, 2003.

[75] H.C. White S.A. Boorman. Social Structure from Multiple Networks. II. Role Structures. *American Journal of Sociology*, 81:1384–1446, 1976.

[76] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[77] A. Salvini. *Analisi delle Reti Sociali: Teorie, Metodi, Applicazioni*. Franco Angeli, Milano, 2007.

[78] J. Scott. *Social Network Analysis: A Handbook*. Sage, London, second edition, 2000.

[79] G. Simmel. *Conflict and the Web of Group Affiliations*. Free Press, Glencoe, IL, 1955.

[80] T.A.B. Snijders. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure*, 3:240, 2002.

[81] L.C. Freeman S.P. Borgatti, M.G. Everett. *UCINET 6 for Windows: Software for Statistical Analysis of of Social Networks*. Harvard.

[82] M.G. Everett S.P. Borgatti. Network Analysis of 2-mode Data. *Social Networks*, 3(19):243–269, August 1997.

[83] J. Spencer. *The Strange Logic of Random Graphs*. Springer, New York, 2000.

[84] D. Strauss. On a General Class of Models for Interaction. *SIAM Rev.*, 28:513–527, 1986.

[85] S.Wasserman. Random Directed Graph Distributions and the Triad Censuses in Social Networks. *Journal of Mathematical Sociology*, 5:61–86, 1977.

[86] S.Wasserman. *Analyzing Social Networks As Stochastic Processes*, volume 75. 1980.

[87] P.E. Pattison S.Wasserman. *Logit Models and Logistic Regressions for Social Networks. I. An Introduction to Markov Graphs and p**, volume 61. 1996.

[88] J. J. Sylvester. On an Application of the New atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics, with Three Gppendices. *American Journal of Mathematic*, 1(1):64–104, March 1878.

[89] K. Thulasiraman. Graphs and vector spaces. In J. L. Gross and J. Yellen, editors, *Handobook of Graph Theory*, Discrete Mathematics and its Applications, chapter 6.

[90] L. Lu W. Aiello, F. Chung. A random graph model for massive graphs. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180, New York, NY, USA, 2000. ACM.

[91] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[92] M. E. Watkins. Automorphisms. In J. L. Gross and J. Yellen, editors, *Handobook of Graph Theory*, Discrete Mathematics and its Applications, chapter 6.

[93] D. J. Watts. Networks, Dynamics, and the Small World Phenomenon. *American Journal of Sociology*, 105:493–592, 1999.

[94] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small World' Networks. *Nature*, 363(6684):409–410, june 1998.

[95] D.J. Watts. *Small Worlds: the Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, NJ, 1999.

[96] H.S. Wilf. *Generatingfunctionology*. Academic Press, New York, second edition, 1994.